

# Mismatched Decoding: Finite-Length Bounds, Error Exponents and Approximations

Jonathan Scarlett, Alfonso Martinez and Albert Guillén i Fàbregas

## Abstract

This paper considers the problem of channel coding with a given (possibly suboptimal) decoding rule. Finite-length upper and lower bounds on the random-coding error probability for a general codeword distribution are given. These bounds are applied to three random-coding ensembles: i.i.d., constant-composition, and cost-constrained. Ensemble-tight error exponents are presented for each ensemble, and achievable second-order coding rates are given. Connections are drawn between the ensembles under both maximum likelihood decoding and mismatched decoding. In particular, it is shown that the error exponents and second-order rates of the constant-composition ensemble can be achieved using cost-constrained coding with at most two cost functions. Finally, saddlepoint approximations of the random-coding bounds are given. These are demonstrated to be more accurate than the approximations obtained from the error exponents and second-order coding rates, while having a similar computational complexity.

## Index Terms

Mismatched decoding, finite-length bounds, random coding, error exponents, second-order coding rates, normal approximation, saddlepoint approximation, maximum-likelihood decoding

J. Scarlett is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: jmscarlett@gmail.com). A. Martinez is with the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: alfonso.martinez@ieee.org). A. Guillén i Fàbregas is with the Institució Catalana de Recerca i Estudis Avançats (ICREA), the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain, and also with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: guillen@ieee.org).

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme (PEOPLE-2011-CIG) under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03. This work was presented in part at the Allerton Conference on Communication, Computing and Control (2012), and at the Information Theory and Applications Workshop (2011, 2013).

This paper was submitted to the IEEE Transactions on Information Theory on March 25th 2013.

## I. INTRODUCTION

It is well known that random coding techniques can be used to prove the achievability part of Shannon's channel coding theorem, as well as characterizing the exponential behavior of the best code for a range of rates under maximum-likelihood (ML) decoding [1, Sec. 5]. In practice, however, ML decoding is often ruled out due to channel uncertainty and implementation constraints. In this paper, we consider the problem of mismatched decoding [2]–[8], in which the decoder chooses the codeword which maximizes a given (possibly suboptimal) decoding metric. In this setting, the following random-coding ensembles have been considered (see Section II for details):

- 1) the i.i.d. ensemble, in which each symbol of each codeword is generated independently;
- 2) the constant-composition ensemble, in which each codeword has the same empirical distribution;
- 3) the cost-constrained ensemble, in which each codeword satisfies a given cost constraint chosen by the codebook designer.

Most existing work has focused on the achievable rates resulting from these ensembles. Motivated by the fact that these rates are strictly smaller than the mismatched capacity in general, the concept of ensemble tightness has been addressed [6]–[8] in order to show that the rates obtained are the best possible for the given random-coding ensemble. The goal of this paper is to present a more comprehensive analysis of the above ensembles, including finite-length bounds and approximations, error exponents, and second-order coding rates.

### A. System Setup

The input and output alphabets are denoted by  $\mathcal{X}$  and  $\mathcal{Y}$  respectively, and the channel is denoted by  $W(y|x)$ . Except where stated otherwise, we assume that  $\mathcal{X}$  and  $\mathcal{Y}$  are finite, and thus the channel is a discrete memoryless channel (DMC). We consider length- $n$  block coding, in which a codebook  $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$  is known at both the encoder and decoder. The encoder takes as input a message  $m$  uniformly distributed on the set  $\{1, \dots, M\}$ , and transmits the corresponding codeword  $\mathbf{x}^{(m)}$ . The decoder receives the vector  $\mathbf{y}$  at the output of the channel, and forms the estimate

$$\hat{m} = \arg \max_{j \in \{1, \dots, M\}} \prod_{i=1}^n q(x_i^{(j)}, y_i), \quad (1)$$

where  $n$  is the length of each codeword and  $x_i^{(j)}$  is the  $i$ -th entry of  $\mathbf{x}^{(j)}$  (similarly for  $y_i$ ). The function  $q(x, y)$  is called the *decoding metric*. We assume that  $q(x, y)$  is greater than or equal to zero, with equality only if  $W(y|x) = 0$ . This assumption entails no real loss of generality, since otherwise the resulting random-coding rates are zero whenever the corresponding input  $x$  is used [3]. In the case of a tie, a random codeword achieving the maximum in (1) is selected. We define  $q^n(\mathbf{x}, \mathbf{y}) \triangleq \prod_{i=1}^n q(x_i, y_i)$  and  $W^n(\mathbf{y}|\mathbf{x}) \triangleq \prod_{i=1}^n W(y_i|x_i)$ .

The mismatched capacity of a given channel and metric is defined to be the supremum of all rates  $R = \frac{1}{n} \log M$  such that the error probability  $p_e(\mathcal{C})$  can be made arbitrarily small for sufficiently large  $n$ . An error exponent  $E(R)$  is said to be achievable if there exists a sequence of codebooks  $\mathcal{C}_n$  of length  $n$  and rate  $R$  such that

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log p_e(\mathcal{C}_n) \geq E(R). \quad (2)$$

We let  $\bar{p}_e(n, M)$  denote the average error probability with respect to a given random-coding ensemble which will be clear from the context. The random-coding error exponent  $E_r(R)$  is said to exhibit ensemble tightness if

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \bar{p}_e(n, e^{nR}) = E_r(R). \quad (3)$$

### B. Previous Work

The most notable early works on mismatched decoding are by Hui [4] and Csisz  r and K  rner [9], who independently derived the achievable rate commonly referred to as the LM rate, given by

$$I^{\text{LM}}(Q) \triangleq \sup_{s \geq 0, a(\cdot)} \mathbb{E} \left[ \log \frac{q(X, Y)^s e^{a(X)}}{\mathbb{E}[q(\bar{X}, Y)^s e^{a(\bar{X})} | Y]} \right], \quad (4)$$

where  $(X, Y) \sim Q \times W$ . This rate can equivalently be expressed as

$$I^{\text{LM}}(Q) = \min_{\tilde{P}_{XY}} I_{\tilde{P}}(X; Y), \quad (5)$$

where the minimization is over all joint distributions on  $\mathcal{X} \times \mathcal{Y}$  satisfying

$$\tilde{P}_X(x) = Q(x) \quad (6)$$

$$\tilde{P}_Y(y) = \sum_x Q(x) W(y|x) \quad (7)$$

$$\sum_{x,y} \tilde{P}_{XY}(x, y) \log q(x, y) \geq \sum_{x,y} Q(x) W(y|x) \log q(x, y). \quad (8)$$

For binary-input DMCs, the LM rate is equal to the mismatched capacity after the optimization over  $Q$  [5]. However, this is not true in general for non-binary channels [3], [6], [8].

The derivation of the LM rate in [4] uses random coding in which the empirical distribution of each codeword is constrained to be close to a given distribution  $Q(x)$ . A similar analysis can be performed for the constant-composition ensemble, yielding the same result [9]. In either case, the proof is valid only when the input and output alphabets are finite. Building on early work by Fischer [10], Kaplan and Shamai [2] performed an analysis which does not rely on the alphabets being finite, and derived the achievable rate known as generalized mutual information (GMI), given by

$$I^{\text{GMI}}(Q) \triangleq \sup_{s \geq 0} \mathbb{E} \left[ \log \frac{q(X, Y)^s}{\mathbb{E}[q(\bar{X}, Y)^s | Y]} \right]. \quad (9)$$

In the discrete memoryless setting, it can be shown that

$$I^{\text{GMI}}(Q) = \min_{\tilde{P}_{XY}} D(\tilde{P}_{XY} \| Q \times \tilde{P}_Y), \quad (10)$$

where the minimization is over all joint distributions on  $\mathcal{X} \times \mathcal{Y}$  satisfying

$$\tilde{P}_Y(y) = \sum_x Q(x) W(y|x) \quad (11)$$

$$\sum_{x,y} \tilde{P}_{XY}(x, y) \log q(x, y) \geq \sum_{x,y} Q(x) W(y|x) \log q(x, y). \quad (12)$$

The GMI cannot exceed the LM rate, and the latter can be strictly higher even after the optimization over  $Q$ . Motivated by this fact, Ganti *et al.* [7] proved that (4) is achievable for general memoryless channels. This was

done using cost-constrained codes, in which each codeword is constrained to satisfy  $\frac{1}{n} \sum_{i=1}^n a(x_i) \approx \mathbb{E}_Q[a(X)]$  for some cost function  $a(\cdot)$  which can be optimized.

It should be noted that the optimization of  $I^{\text{LM}}(Q)$  and  $I^{\text{GMI}}(Q)$  over  $Q$  is questionable in some applications, since it requires the codebook designer to know the channel. For the cost-constrained random-coding ensemble, a similar statement is true for the optimization over  $a(\cdot)$  in (4). On the other hand, mismatched decoding is relevant in settings where the channel is known but implementation constraints prohibit the use of the ML decoder, and such optimizations are valid in these settings. Throughout this paper, we present results for a given  $Q$ . Most results presented for the cost-constrained ensemble will assume a fixed set of parameters, but we will sometimes choose  $a(\cdot)$  as a function of the channel. In such cases, it is assumed that the codebook designer knows the channel.

In the terminology of [7], (5) and (10) are primal expressions, and (4) and (9) are the corresponding dual expressions. Indeed, the latter can be derived from the former using Lagrange duality techniques [11]. As well as extending readily to general alphabets, the dual expressions have the advantage that an achievable rate can be obtained by substituting any values of  $s$  and  $a(\cdot)$  into (4) and (9), whereas (5) and (10) only give valid bounds after the minimization is performed.

Due to the lack of converse results in mismatched decoding, it is of interest to determine whether the achievable rates are tight with respect to the ensemble average. For the i.i.d. ensemble with input distribution  $Q$ , it is known that the random-coding error probability  $\bar{p}_e(n, e^{nR})$  tends to 0 as  $n \rightarrow \infty$  when  $R < I^{\text{GMI}}(Q)$ , whereas  $\bar{p}_e(n, e^{nR}) \rightarrow 1$  as  $n \rightarrow \infty$  when  $R > I^{\text{GMI}}(Q)$  [2], [7]. Similarly, for the constant-composition ensemble with input distribution  $Q$ , it has been shown that  $\bar{p}_e(n, e^{nR}) \rightarrow 0$  as  $n \rightarrow \infty$  when  $R < I^{\text{LM}}(Q)$ , whereas  $\bar{p}_e(n, e^{nR}) \rightarrow 1$  as  $n \rightarrow \infty$  when  $R > I^{\text{LM}}(Q)$  [6], [8].

Finally, random-coding error exponents under mismatched decoding are presented in [2], [9], [12]. The exponent in [9] is valid for any DMC, and is proved using constant-composition random coding. Error exponents for the i.i.d. ensemble and cost-constrained ensemble are presented in [2] and [12] respectively, and both results apply in the case of general alphabets. The exponents in [9], [12] recover the LM rate, whereas the exponent in [2] only recovers the GMI. To our knowledge, the ensemble tightness of the exponent has been shown only in the case of constant-composition random coding with a sufficiently small rate [13].

### C. Contributions

Motivated by the fact that most existing work on mismatched decoding has focused on achievable rates, the main goal of this paper is to present a more detailed analysis of the random-coding error probability. Our main contributions are as follows:

- 1) Finite-length upper and lower bounds on the random-coding error probability for a general random-coding distribution are given. These bounds are applied to three random-coding ensembles: the i.i.d. ensemble, constant-composition ensemble and cost-constrained ensemble.
- 2) Ensemble-tight error exponents are given for each ensemble, and connections are drawn between the ensembles.

For the i.i.d. and constant-composition ensembles, our analysis proves the ensemble tightness of the exponents

given in [2] and [9] respectively. The exponent for the cost-constrained ensemble appears to be new.

- 3) Achievable second-order coding rates are presented for each ensemble. We focus on the cost-constrained ensemble, for which we derive the second-order rate using novel techniques.
- 4) The benefits of cost-constrained random-coding with multiple cost functions are studied. In particular, while the constant-composition ensemble gives the best exponents and second-order rates of the ensembles considered, it is shown that the same performance can be achieved using cost-constrained coding with two cost functions. The latter does not require the input alphabet to be finite, and thus the results hold in greater generality.
- 5) Saddlepoint approximations of certain random-coding bounds are presented for the i.i.d. and constant-composition ensembles. These are shown to provide accurate approximations to the corresponding bounds at all rates, even at small block lengths. In contrast, the approximations obtained from the error exponents and second-order rates are generally accurate only at large block lengths, or only for a narrower range of rates. For the i.i.d. ensemble, we prove that the saddlepoint approximation yields the same behavior as the corresponding random-coding bound as  $n \rightarrow \infty$  with a fixed rate, to within a multiplicative  $1 + o(1)$  term.

#### D. Notation

The set of all probability distributions on an alphabet  $\mathcal{A}$  is denoted by  $\mathcal{P}(\mathcal{A})$ , and the set of all empirical distributions on a vector in  $\mathcal{A}^n$  (i.e. types [14, Sec. 2] [15]) is denoted by  $\mathcal{P}_n(\mathcal{A})$ . The type of a vector  $\mathbf{x}$  is denoted by  $\hat{P}_{\mathbf{x}}(\cdot)$ . For a given  $Q \in \mathcal{P}_n(\mathcal{A})$ , the type class  $T^n(Q)$  is defined to be the set of all sequences in  $\mathcal{A}^n$  with type  $Q$ .

The probability of an event is denoted by  $\mathbb{P}[\cdot]$ , and the symbol  $\sim$  means “distributed as”. The marginals of a joint distribution  $P_{XY}(x, y)$  are denoted by  $P_X(x)$  and  $P_Y(y)$ . We write  $P_X = \tilde{P}_X$  to denote element-wise equality between two probability distributions on the same alphabet. Expectation with respect to a joint distribution  $P_{XY}(x, y)$  is denoted by  $\mathbb{E}_P[\cdot]$ . When the associated probability distribution is understood from the context, the expectation is written as  $\mathbb{E}[\cdot]$ . Similarly, mutual information with respect to  $P_{XY}$  is written as  $I_P(X; Y)$ , or simply  $I(X; Y)$  when the distribution is understood from the context. Given a distribution  $Q(x)$  and conditional distribution  $W(y|x)$ , we write  $Q \times W$  to denote the joint distribution defined by  $Q(x)W(y|x)$ . The set of values where a probability mass function  $Q$  is non-zero is denoted by  $\text{supp}(Q)$ .

For two sequences  $f(n)$  and  $g(n)$ , we write  $f(n) \doteq g(n)$  if  $\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{f(n)}{g(n)} = 0$ , and similarly for  $\dot{\leq}$  and  $\dot{\geq}$ . We write  $f(n) = O(g(n))$  if  $|f(n)| \leq c|g(n)|$  for some  $c$  and sufficiently large  $n$ . We write  $f(n) = \Omega(g(n))$  if  $g(n) = O(f(n))$ ,  $f(n) = \Theta(g(n))$  if both  $f(n) = O(g(n))$  and  $f(n) = \Omega(g(n))$  hold, and  $f(n) = o(g(n))$  if  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$ .

We denote the tail probability of a zero-mean unit-variance Gaussian variable by  $Q(\cdot)$ , and we denote its functional inverse by  $Q^{-1}(\cdot)$ . The complementary error function is denoted by  $\text{erfc}(\cdot)$ , i.e.  $\text{erfc}(z) = 2Q(\sqrt{2}z)$ . All logarithms have base  $e$ , and all rates are in units of nats except in the examples, where bits are used. We define  $[c]^+ = \max\{0, c\}$ , and denote the indicator function by  $\mathbb{1}\{\cdot\}$ .

### E. Structure of the Paper

In Section II, we formally define the random-coding ensembles used throughout the paper. In Section III, we give upper and lower bounds on the random-coding error probability. In Section IV, we derive ensemble-tight error exponents for each ensemble, and draw connections between the ensembles. Second-order coding rates are presented in Section V, and saddlepoint approximations are presented in Section VI. Conclusions are drawn in Section VII.

## II. RANDOM-CODING ENSEMBLES

In this section, we formally define the three random-coding ensembles described in Section I. Throughout the paper, we denote the random-coding distribution by  $P_{\mathbf{X}}$ , and we assume that each codeword is generated independently. We pay particular attention to the following.

- 1) The i.i.d. ensemble is characterized by

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n Q(x_i). \quad (13)$$

In words, each symbol of each codeword is generated independently according to  $Q$ .

- 2) The constant-composition ensemble is characterized by

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|T^n(Q_n)|} \mathbb{1}\{\mathbf{x} \in T^n(Q_n)\}, \quad (14)$$

where  $Q_n$  is a type such that  $\max_x |Q_n(x) - Q(x)| \leq \frac{1}{n}$ . That is, each codeword is generated uniformly over the type class  $T^n(Q_n)$ , and hence each codeword has the same composition.

- 3) The cost-constrained ensemble is characterized by

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\mu_n} \prod_{i=1}^n Q(x_i) \mathbb{1}\{\mathbf{x} \in \mathcal{D}_n\}, \quad (15)$$

where

$$\mathcal{D}_n \triangleq \left\{ \mathbf{x} : \left| \frac{1}{n} \sum_{i=1}^n a_l(x_i) - \phi_l \right| \leq \frac{\delta}{n}, l = 1, \dots, L \right\}, \quad (16)$$

and where  $\mu_n$  is a normalizing constant,  $\delta$  is a positive constant (independent of  $n$ ), and for each  $l \in \{1, \dots, L\}$ ,  $a_l(\cdot)$  is a cost function and  $\phi_l \triangleq \mathbb{E}_Q[a_l(X)]$ . Roughly speaking, each codeword is generated according to an i.i.d. distribution conditioned on the empirical mean of each cost function  $a_l(x)$  being close to the true mean.

This generalizes the ensemble studied in [1, Sec. 7.3] [12] by including multiple costs.

The cost functions  $\{a_l\}$  in (15) should not be viewed as being chosen to meet a system constraint (e.g. power limitations). Rather, they are introduced in order to improve the performance of the random-coding ensemble. Thus, one can view the costs as being *pseudo-costs*, rather than system costs. However, the latter can be handled similarly; see Section VII for further discussion.

The constant  $\delta$  in (15) could, in principle, vary with  $l$  and  $n$ . However, a fixed value will suffice for our purposes. The constant-composition ensemble is a special case of the ensemble in (15), since it is obtained by replacing  $Q$  by  $Q_n$  (see (14)), setting  $L = |\mathcal{X}| - 1$ , and choosing the cost functions  $a_1 = (1, 0, \dots, 0)$ ,  $a_2 = (0, 1, 0, \dots, 0)$ ,

etc., along with  $\delta < 1$ . Of course, the i.i.d. ensemble is also a special case of (15), since it is obtained by setting  $L = 0$ .

The following proposition shows that the normalizing constant  $\mu_n$  in (15) decays at most polynomially in  $n$ . When  $|\mathcal{X}|$  is finite, this can easily be shown using the method of types. In particular, choosing the functions given in the previous paragraph to recover the constant-composition ensemble, we have  $\mu_n \geq (n+1)^{-(|\mathcal{X}|-1)}$  [14, pp. 17]. For the sake of generality, we present a proof which applies to more general alphabets, subject to minor technical conditions. The proof for the case  $L = 1$  is given in [1, Ch. 7.3].

**Proposition 1.** *Fix an input alphabet  $\mathcal{X}$  (possibly infinite or continuous), an input distribution  $Q \in \mathcal{P}(\mathcal{X})$  and the cost functions  $a_1(\cdot), \dots, a_L(\cdot)$ . If the second moment of  $a_l(X)$  is finite under  $X \sim Q$  for  $l = 1, \dots, L$ , then there exists a  $\delta > 0$  such that the normalizing constant in (15) satisfies  $\mu_n = \Omega(n^{-L/2})$ .*

*Proof:* Defining the vectors  $\phi = (\phi_1, \dots, \phi_L)^T$  and  $\mathbf{a}(x) = (a_1(x), \dots, a_L(x))^T$ , we have

$$\mu_n = \mathbb{P} \left[ \frac{\delta}{\sqrt{n}} \mathbf{1} \preceq \frac{1}{\sqrt{n}} \left( \sum_{i=1}^n \mathbf{a}(X_i) - n\phi \right) \preceq \frac{\delta}{\sqrt{n}} \mathbf{1} \right], \quad (17)$$

where  $\preceq$  denotes element-wise inequality, and  $\mathbf{1}$  is the vector of ones. We define  $\Sigma$  to be the covariance matrix of  $\mathbf{a}(X) - \phi$ . Since each value of  $\phi_l$  and  $\sigma_l^2 \triangleq \mathbb{E}[(a_l(X) - \phi_l)^2]$  is finite by assumption, we can apply the multivariate central limit theorem [16, Sec. VIII], yielding that the distribution of  $\frac{1}{\sqrt{n}} \left( \sum_{i=1}^n \mathbf{a}(X_i) - n\phi \right)$  converges to that of multivariate normal random variable with zero mean and covariance matrix  $\Sigma$ .

We will show that  $\mu_n = \Theta(n^{-L/2})$  provided that  $\det(\Sigma) > 0$ . In the case that  $\det(\Sigma) = 0$ , the evaluation of  $\mu_n$  can be reduced to a lower dimension  $L'$ , yielding  $\mu_n = \Theta(n^{-L'/2})$  and thus proving the general result  $\mu_n = \Omega(n^{-L/2})$ . The proof of the case  $\det(\Sigma) > 0$  is most easily understood in the case that the density of  $\frac{1}{\sqrt{n}} \left( \sum_{i=1}^n \mathbf{a}(X_i) - n\phi \right)$  converges to that of a multivariate normal random variable, namely

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{L/2} \det(\Sigma)^{1/2}} \exp \left( -\frac{1}{2} \mathbf{z}^T \Sigma^{-1} \mathbf{z} \right). \quad (18)$$

Such convergence is not obtained in general, since convergence in distribution does not imply convergence of the density [16, Sec. XV].

Since the density in (18) approaches  $\frac{1}{(2\pi)^{L/2} \det(\Sigma)^{1/2}}$  as each entry of  $\mathbf{z}$  tends to zero, and since the event in (17) constrains each entry to be  $\frac{\delta}{\sqrt{n}}$ -close to its mean, we have

$$\mu_n = \frac{\left( \frac{2\delta}{\sqrt{n}} (1 + o(1)) \right)^L}{(2\pi)^{L/2} \det(\Sigma)^{1/2}} \quad (19)$$

thus yielding  $\mu_n = \Theta(n^{-L/2})$  as desired. To handle the general case, we make use of [17, Thm. 6.4], which gives asymptotic expressions for probabilities of sets of the form (16). In particular, the desired result  $\mu_n = \Theta(n^{-L/2})$  follows by choosing  $\delta$  to be greater than or equal to the largest span of the  $a_l(X)$  which are lattice random variables. In the case that all of the  $a_l(X)$  are non-lattice random variables,  $\delta$  can be chosen arbitrarily. ■

In accordance with Proposition 1, we henceforth assume that the choice of  $\delta$  for the cost-constrained ensemble is such that  $\mu_n = \Omega(n^{-L/2})$ .

### III. RANDOM-CODING ERROR PROBABILITY

In recent years, the problem of characterizing the finite-length performance of coded communication systems has regained significant attention. In particular, several upper and lower bounds on the smallest error probability for a given block length  $n$  and number of messages  $M$  were provided by Polyanskiy *et al.* [18]. In this section, we obtain finite-length upper bounds on the error probability in the mismatched setting. Furthermore, we address the problem of ensemble tightness in the finite-length regime by obtaining lower bounds to the random-coding error probability.

#### A. Finite-Length Bounds

Throughout this section, we make use of the random variables

$$(\mathbf{X}, \mathbf{Y}, \bar{\mathbf{X}}) \sim P_{\mathbf{X}}(\mathbf{x})W^n(\mathbf{y}|\mathbf{x})P_{\mathbf{X}}(\bar{\mathbf{x}}). \quad (20)$$

As stated in [18], the exact random-coding error probability of the decoder that breaks ties randomly is given by

$$\bar{p}_e(n, M) = 1 - \sum_{\ell=0}^{M-1} \binom{M-1}{\ell} \frac{1}{\ell+1} \mathbb{E} \left[ p_0(\mathbf{X}, \mathbf{Y})^\ell p_1(\mathbf{X}, \mathbf{Y})^{M-1-\ell} \right], \quad (21)$$

where

$$p_0(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{y}) = q^n(\mathbf{x}, \mathbf{y})] \quad (22)$$

$$p_1(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{y}) < q^n(\mathbf{x}, \mathbf{y})]. \quad (23)$$

The computation of (21) is generally infeasible even for small values of  $n$ . However, we can weaken (21) to obtain computationally tractable bounds. The following theorem extends the RCU bound of [18] to the mismatched setting, and provides a weakened bound with a similar form.

**Theorem 1.** (Upper Bounds) *For any random-coding distribution  $P_{\mathbf{X}}(\mathbf{x})$ , the random-coding error probability satisfies*

$$\bar{p}_e(n, M) \leq \text{rcu}(n, M) \leq \text{rcu}_s(n, M), \quad (24)$$

where

$$\text{rcu}(n, M) \triangleq \mathbb{E} \left[ \min \left\{ 1, (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y}) \mid \mathbf{X}, \mathbf{Y}] \right\} \right] \quad (25)$$

$$\text{rcu}_s(n, M) \triangleq \mathbb{E} \left[ \min \left\{ 1, (M-1) \frac{\mathbb{E}[q^n(\bar{\mathbf{X}}, \mathbf{Y})^s \mid \mathbf{Y}]}{q^n(\mathbf{X}, \mathbf{Y})^s} \right\} \right] \quad (26)$$

and (24) holds for any  $s \geq 0$ .

*Proof:* Similarly to [18], (25) follows by upper bounding the error probability by that of the decoder which decodes ties as errors, and then applying the truncated union bound to

$$\mathbb{P} \left[ \bigcup_{i \neq m} \{q^n(\mathbf{X}^{(i)}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y})\} \right] \quad (27)$$



after conditioning on  $\mathbf{X}$  and  $\mathbf{Y}$ . We obtain (26) using Markov's inequality.  $\blacksquare$

As discussed in Section I, it is of interest to find lower bounds on the random-coding error probability in the mismatched setting. This is addressed in the following theorem.

**Theorem 2.** (Lower Bounds) *The random-coding error probability for the mismatched decoder which resolves ties randomly satisfies*

$$\bar{p}_e(n, M) \geq \text{rcu}_L(n, M) \geq \text{rcu}'_L(n, M) \geq \frac{1}{2} \left(1 - \frac{1}{e}\right) \text{rcu}(n, M), \quad (28)$$

where

$$\text{rcu}_L(n, M) \triangleq 1 - \frac{1}{2} \mathbb{E} \left[ \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) < q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}]^{M-1} \right] - \frac{1}{2} \mathbb{E} \left[ \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \leq q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}]^{M-1} \right] \quad (29)$$

$$\begin{aligned} \text{rcu}'_L(n, M) \triangleq 1 - \frac{1}{2} \mathbb{E} \left[ \exp \left( - (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}] \right) \right] \\ - \frac{1}{2} \mathbb{E} \left[ \exp \left( - (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) > q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}] \right) \right]. \end{aligned} \quad (30)$$

*Proof:* Consider a fixed codebook  $\mathcal{C}$  with error probability  $p_e(\mathcal{C})$ . Let  $B_0$  be the event that one or more codewords yield a strictly higher metric than the transmitted one, and let  $B_\ell$  ( $\ell \geq 1$ ) be the event that the transmitted codeword yields a metric which is equal highest with  $\ell$  other codewords. We have

$$p_e(\mathcal{C}) = \mathbb{P}[B_0] + \sum_{\ell=1}^{M-1} \mathbb{P}[B_\ell] \frac{\ell}{\ell+1} \quad (31)$$

$$\geq \mathbb{P}[B_0] + \frac{1}{2} \sum_{\ell=1}^{M-1} \mathbb{P}[B_\ell] \quad (32)$$

$$= \frac{1}{2} p'_e(\mathcal{C}) + \frac{1}{2} \mathbb{P}[B_0], \quad (33)$$

where  $p'_e(\mathcal{C}) \triangleq \mathbb{P}[B_0] + \sum_{\ell=1}^{M-1} \mathbb{P}[B_\ell]$  is the error probability of the mismatched decoder which decodes ties as errors. Averaging (33) over the random-coding distribution, we obtain

$$\bar{p}_e(n, M) \geq \frac{1}{2} \mathbb{P} \left[ \bigcup_{i=2}^M \{q^n(\mathbf{X}^{(i)}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y})\} \right] + \frac{1}{2} \mathbb{P} \left[ \bigcup_{i=2}^M \{q^n(\mathbf{X}^{(i)}, \mathbf{Y}) > q^n(\mathbf{X}, \mathbf{Y})\} \right]. \quad (34)$$

The first inequality in (28) follows by writing each probability in (34) as an expectation given  $\mathbf{X}$  and  $\mathbf{Y}$ , and using the fact that for a sequence of independent events  $A_1, \dots, A_k$  having equal probability,

$$\mathbb{P} \left[ \bigcup_{i=1}^k A_i \right] = 1 - \mathbb{P} \left[ \bigcap_{i=1}^k A_i^c \right] \quad (35)$$

$$= 1 - (1 - \mathbb{P}[A_1])^k. \quad (36)$$

The second inequality in (28) follows since  $1 - \alpha \leq e^{-\alpha}$ . The final inequality in (28) follows by lower bounding

(30) as follows:

$$\text{rcu}'_{\text{L}}(n, M) \geq \frac{1}{2} \left( 1 - \mathbb{E} \left[ \exp \left( - (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}] \right) \right] \right) \quad (37)$$

$$\geq \frac{1}{2} \left( 1 - \mathbb{E} \left[ \exp \left( - \min \left\{ 1, (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}] \right\} \right) \right] \right) \quad (38)$$

$$\geq \frac{1}{2} \left( 1 - \frac{1}{e} \right) \mathbb{E} \left[ \min \left\{ 1, (M-1) \mathbb{P}[q^n(\bar{\mathbf{X}}, \mathbf{Y}) \geq q^n(\mathbf{X}, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}] \right\} \right], \quad (39)$$

where (39) follows since  $e^{-\alpha} \geq (1 - \frac{1}{e})\alpha$  for  $\alpha \in [0, 1]$ .  $\blacksquare$

We refer to  $\text{rcu}(n, M)$ ,  $\text{rcu}_s(n, M)$ ,  $\text{rcu}_{\text{L}}(n, M)$  and  $\text{rcu}'_{\text{L}}(n, M)$  as random-coding union (RCU) bounds [18]. The bounds  $\text{rcu}_{\text{L}}$  and  $\text{rcu}'_{\text{L}}$  have a similar form; the latter is slightly weaker but more numerically stable. From (28), we see that  $\text{rcu}$  and  $\text{rcu}_{\text{L}}$  coincide to within the constant factor  $\frac{1}{2}(1 - \frac{1}{e})$ , and thus accurately describe the random-coding error probability. However, for general channels, decoding metrics and random-coding distributions, their computation can be challenging. Generally, some degree of symmetry is needed in order to facilitate the computation. Since constant factors do not affect the error exponent, we have  $\text{rcu} \doteq \text{rcu}_{\text{L}} \doteq \bar{p}_e$ . In Section IV, this observation will be used to obtain ensemble-tight error exponents for the ensembles introduced in Section II.

The bounds  $\text{rcu}$  and  $\text{rcu}_s$  generally have a similar computational complexity. However, we will see in Section VI that  $\text{rcu}_s$  can be approximated accurately using the saddlepoint approximation. In particular, for the i.i.d. ensemble, we will obtain an approximation  $\widehat{\text{rcu}}_s$  with low computational complexity such that  $\lim_{n \rightarrow \infty} \frac{\widehat{\text{rcu}}_s(n, e^{nR})}{\text{rcu}_s(n, e^{nR})} = 1$  for rates  $R$  below the GMI, subject to minor technical conditions. Furthermore, we will use  $\text{rcu}_s$  to derive achievable error exponents in Section IV, and achievable second-order rates in Section V.

The focus in this paper is on memoryless channels with single-letter decoding metrics and the random-coding ensembles given in Section II. However, the above bounds apply to channels with memory, multi-letter decoding rules and general random-coding ensembles with independently generated codewords each having the same distribution. In fact, the independence assumption is not needed for the upper bounds, and lower bounds in the *pairwise* independent case can be obtained by replacing (36) with de Caen's lower bound [19]; see [20, Theorem 1].

### B. Numerical Results

We compare the upper and lower bounds numerically by considering the channel defined by the entries of the  $|\mathcal{X}| \times |\mathcal{Y}|$  matrix

$$\begin{bmatrix} 1 - 2\delta_0 & \delta_0 & \delta_0 \\ \delta_1 & 1 - 2\delta_1 & \delta_1 \\ \delta_2 & \delta_2 & 1 - 2\delta_2 \end{bmatrix} \quad (40)$$

with  $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ . The mismatched decoder chooses the codeword which is closest to  $\mathbf{y}$  in terms of Hamming distance. For example, the decoding metric can be taken to be the entries of the matrix

$$\begin{bmatrix} 1 - 2\delta & \delta & \delta \\ \delta & 1 - 2\delta & \delta \\ \delta & \delta & 1 - 2\delta \end{bmatrix} \quad (41)$$

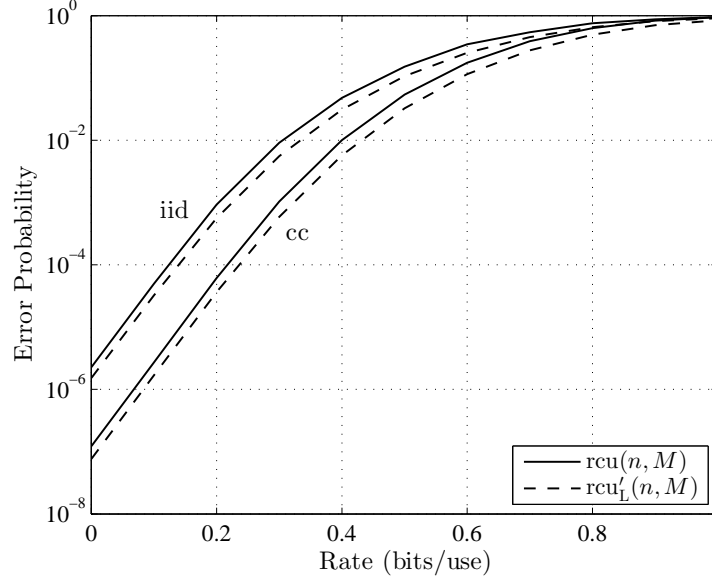


Figure 1. Upper and lower bounds on the random-coding error probability for the channel defined in (40) with  $n = 45$ ,  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$  and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . The mismatched decoder uses the minimum Hamming distance metric given in (41).

for any  $\delta \in (0, \frac{1}{3})$ . That is, the channel is asymmetric, but the decoder uses a symmetric decoding metric.

We first set  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$  and  $\delta_2 = 0.25$  and consider the i.i.d. and constant-composition ensembles with  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ; the cost-constrained ensemble is omitted because the multi-letter nature of the ensemble makes it difficult to compute the finite-length bounds. Under these parameters we have  $I^{\text{GM}}(Q) = 0.643$ ,  $I^{\text{LM}}(Q) = 0.728$  and  $I(X; Y) = 0.763$  bits/use. Figure 1 plots  $\text{rcu}(n, M)$  and  $\text{rcu}'_L(n, M)$  with  $n = 45$ . We confirm the close match between the upper and lower bounds predicted by Theorem 2. It should be noted that this gap closes further when the decoding metric yields fewer ties [20, Fig. 1]. The present metric yields a significant number of ties due to the fact that  $q(x, y)$  can only be one of two different values, namely  $\delta$  and  $1 - 2\delta$ .

We see that the constant-composition ensemble yields a lower error probability than that of the i.i.d. ensemble, particularly at low to moderate rates. This shows that the performance gain of the constant-composition ensemble can be significant even at small block lengths.

To facilitate the computation at larger block lengths, we consider the symmetric setup of  $\delta_0 = \delta_1 = \delta_2 = \delta = 0.1$ , and focus on the i.i.d. ensemble with  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . In this case, the decoder is simply the ML decoder, and the mutual information is  $I(X; Y) = 0.633$  bits/use. Figure 2 plots the bounds of Theorems 1 and 2 with  $n = 250$ . For the bound  $\text{rcu}_s$ , we choose  $s$  to maximize the random-coding error exponent at each rate  $R$ ; see Section IV for details. Once again,  $\text{rcu}(n, M)$  and  $\text{rcu}'_L(n, M)$  characterize the random-coding error probability accurately over the entire range plotted. The bound  $\text{rcu}_s(n, M)$  is only slightly looser than  $\text{rcu}(n, M)$ . The gap between the two can be larger at smaller block lengths, and is characterized by a multiplicative  $O(\frac{1}{\sqrt{n}})$  term for sufficiently regular channels and metrics; see Section VI for further discussion.

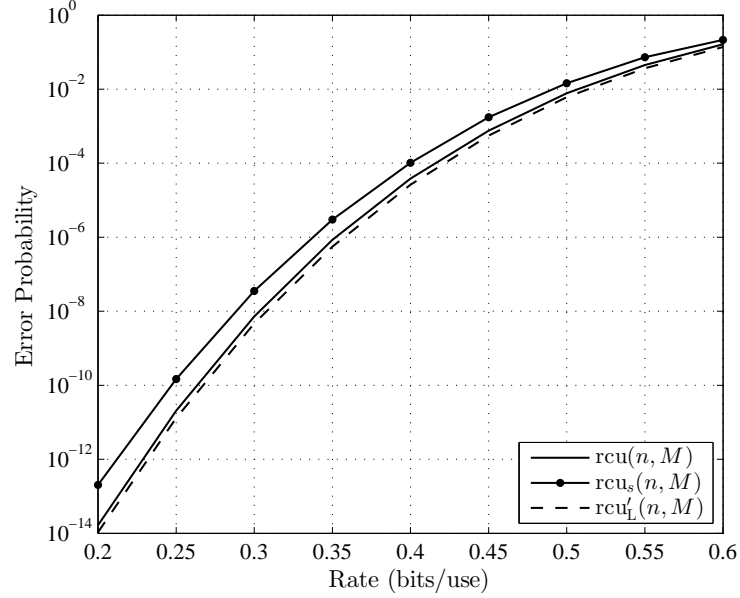


Figure 2. Upper and lower bounds on the random-coding error probability for the channel defined in (40) with  $n = 250$ ,  $\delta_0 = \delta_1 = \delta_2 = 0.1$ ,  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  and ML decoding.

#### IV. RANDOM-CODING ERROR EXPONENT

Error exponents characterize the asymptotic exponential behavior of the error probability in coded communication systems, and thus provide more information than capacity results alone. In the matched setting, error exponents were studied by Fano [21, Sec. 9] and later by Gallager *et al.* [1, Sec. 5]. The ensemble tightness of the exponent (cf. (3)) under ML decoding was studied by Gallager [22] and D'yachkov [23] for the i.i.d. and constant-composition ensembles respectively.

In this section, we present a unified analysis for obtaining the error exponents of the random-coding ensembles given in Section II in the mismatched setting. We prove the ensemble tightness of each exponent, and provide several connections between the ensembles.

We define the sets

$$\mathcal{S}^{\text{iid}} \triangleq \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \quad (42)$$

$$\mathcal{S}^{\text{cc}}(Q) \triangleq \{P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : P_X = Q\} \quad (43)$$

$$\mathcal{S}^{\text{cost}}(\{a_l\}) \triangleq \{P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \mathbb{E}_P[a_l(X)] = \phi_l, l = 1, \dots, L\} \quad (44)$$

and

$$\mathcal{T}^{\text{iid}}(P_{XY}) \triangleq \left\{ \tilde{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \tilde{P}_Y = P_Y, \mathbb{E}_{\tilde{P}}[\log q(X, Y)] \geq \mathbb{E}_P[\log q(X, Y)] \right\} \quad (45)$$

$$\mathcal{T}^{\text{cc}}(P_{XY}) \triangleq \left\{ \tilde{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \tilde{P}_X = P_X, \tilde{P}_Y = P_Y, \mathbb{E}_{\tilde{P}}[\log q(X, Y)] \geq \mathbb{E}_P[\log q(X, Y)] \right\} \quad (46)$$

$$\mathcal{T}^{\text{cost}}(P_{XY}, \{a_l\}) \triangleq \left\{ \tilde{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \mathbb{E}_{\tilde{P}}[a_l(X)] = \phi_l, l = 1, \dots, L, \right. \\ \left. \tilde{P}_Y = P_Y, \mathbb{E}_{\tilde{P}}[\log q(X, Y)] \geq \mathbb{E}_P[\log q(X, Y)] \right\}, \quad (47)$$

where the notation  $\{a_l\}$  is used to denote dependence on  $a_1(\cdot), \dots, a_L(\cdot)$ .

**Theorem 3.** *The ensemble-tight random-coding error exponents for the ensembles defined in (13)–(15) are respectively given by*

$$E_r^{\text{iid}}(Q, R) \triangleq \min_{P_{XY} \in \mathcal{S}^{\text{iid}}} \min_{\tilde{P}_{XY} \in \mathcal{T}^{\text{iid}}(P_{XY})} D(P_{XY} \| Q \times W) + \left[ D(\tilde{P}_{XY} \| Q \times \tilde{P}_Y) - R \right]^+ \quad (48)$$

$$E_r^{\text{cc}}(Q, R) \triangleq \min_{P_{XY} \in \mathcal{S}^{\text{cc}}(Q)} \min_{\tilde{P}_{XY} \in \mathcal{T}^{\text{cc}}(P_{XY})} D(P_{XY} \| Q \times W) + \left[ I_{\tilde{P}}(X; Y) - R \right]^+ \quad (49)$$

$$E_r^{\text{cost}}(Q, R, \{a_l\}) \triangleq \min_{P_{XY} \in \mathcal{S}^{\text{cost}}(\{a_l\})} \min_{\tilde{P}_{XY} \in \mathcal{T}^{\text{cost}}(P_{XY}, \{a_l\})} D(P_{XY} \| Q \times W) + \left[ D(\tilde{P}_{XY} \| Q \times \tilde{P}_Y) - R \right]^+. \quad (50)$$

*Proof:* See Appendix A. ■

The optimization problems in (48)–(50) are all convex when the input distribution and cost functions are fixed. Using the method of Lagrange duality [11], each exponent can be written in an alternative form.

**Theorem 4.** *The error exponents in (48)–(50) can be expressed as*

$$E_r^{\text{iid}}(Q, R) = \max_{\rho \in [0, 1]} E_0^{\text{iid}}(Q, \rho) - \rho R \quad (51)$$

$$E_r^{\text{cc}}(Q, R) = \max_{\rho \in [0, 1]} E_0^{\text{cc}}(Q, \rho) - \rho R \quad (52)$$

$$E_r^{\text{cost}}(Q, R, \{a_l\}) = \max_{\rho \in [0, 1]} E_0^{\text{cost}}(Q, \rho, \{a_l\}) - \rho R, \quad (53)$$

where

$$E_0^{\text{iid}}(Q, \rho) \triangleq \sup_{s \geq 0} -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s | Y]}{q(X, Y)^s} \right)^\rho \right] \quad (54)$$

$$E_0^{\text{cc}}(Q, \rho) \triangleq \sup_{s \geq 0, a(\cdot)} \mathbb{E} \left[ -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{a(\bar{X})} | Y]}{q(X, Y)^s e^{a(X)}} \right)^\rho \middle| X \right] \right] \quad (55)$$

$$E_0^{\text{cost}}(Q, \rho, \{a_l\}) \triangleq \sup_{s \geq 0, \{r_l\}, \{\bar{r}_l\}} -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{\sum_{l=1}^L \bar{r}_l(a_l(\bar{X}) - \phi_l)} | Y]}{q(X, Y)^s e^{\sum_{l=1}^L r_l(a_l(X) - \phi_l)}} \right)^\rho \right] \quad (56)$$

and  $(X, Y, \bar{X}) \sim Q(x)W(y|x)Q(\bar{x})$ .

*Proof:* The proofs are similar for each of the three ensembles; see Appendix B for the proof of (52). ■

We refer to (48)–(50) as primal expressions for the exponents, and (51)–(53) as dual expressions. The exponents in (49) and (51) were derived in [9] and [2] respectively, though both derivations were different to ours. To our

knowledge, the alternative expressions in (48) and (52) have not appeared previously, and the exponent  $E_r^{\text{cost}}$  is new.

The exponents in (51)–(53) can be derived directly, rather than via Lagrange duality. The derivation of (51) is presented in [2], and the expression in (52) follows by combining the achievable error exponent of [24] with the fact that under constant composition codes, the metric  $q(x, y)$  is equivalent to the metric  $q(x, y)^s e^{a(x)}$  for any  $s \geq 0$  and  $a(\cdot)$  [3].

The direct derivation of (53) can be considered as a refinement of that of [12], where it was shown that an achievable error exponent in the case that  $L = 1$  is given by

$$E_r^{\text{cost}'}(Q, R, a_1) \triangleq \max_{\rho \in [0, 1]} E_0^{\text{cost}'}(Q, \rho, a_1) - \rho R \quad (57)$$

$$E_0^{\text{cost}'}(Q, \rho, a_1) \triangleq \sup_{s \geq 0} -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{a_1(\bar{X})} | Y]}{q(X, Y)^s e^{a_1(X)}} \right)^\rho \right]. \quad (58)$$

For completeness, we include the derivation of  $E_r^{\text{cost}}$  here. Applying  $\min\{1, \alpha\} \leq \alpha^\rho$  ( $\rho \in [0, 1]$ ) to  $\text{rcu}_s$  in (26), we obtain

$$\bar{p}_e(n, M) \leq \frac{1}{\mu_n^{1+\rho}} M^\rho \sum_{\mathbf{x} \in \mathcal{D}_n, \mathbf{y}} Q^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}} \in \mathcal{D}_n} Q^n(\bar{\mathbf{x}}) q^n(\bar{\mathbf{x}}, \mathbf{y})^s}{q^n(\mathbf{x}, \mathbf{y})^s} \right)^\rho, \quad (59)$$

where  $Q^n(\mathbf{x}) \triangleq \prod_{i=1}^n Q(x_i)$ , and  $\mathcal{D}_n$  is defined in (16). From the definition of  $\mathcal{D}_n$ , each codeword  $\mathbf{x} \in \mathcal{D}_n$  satisfies

$$\exp(r(a_l^n(\mathbf{x}) - n\phi_l)) \leq e^{|r|\delta} \quad (60)$$

for any real number  $r$ , where  $a_l^n(\mathbf{x}) \triangleq \sum_{i=1}^n a_l(x_i)$ . Weakening (59) by applying (60) multiple times, we obtain

$$\bar{p}_e(n, M) \leq \frac{e^{\rho \sum_l (|r_l| + |\bar{r}_l|)\delta}}{\mu_n^{1+\rho}} M^\rho \sum_{\mathbf{x} \in \mathcal{D}_n, \mathbf{y}} Q^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}} \in \mathcal{D}_n} Q^n(\bar{\mathbf{x}}) q^n(\bar{\mathbf{x}}, \mathbf{y})^s e^{\sum_l r_l (a_l^n(\bar{\mathbf{x}}) - n\phi_l)}}{q^n(\mathbf{x}, \mathbf{y})^s e^{\sum_l \bar{r}_l (a_l^n(\mathbf{x}) - n\phi_l)}} \right)^\rho \quad (61)$$

$$\leq \frac{e^{\rho \sum_l (|r_l| + |\bar{r}_l|)\delta}}{\mu_n^{1+\rho}} M^\rho \sum_{\mathbf{x}, \mathbf{y}} Q^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}}) q^n(\bar{\mathbf{x}}, \mathbf{y})^s e^{\sum_l r_l (a_l^n(\bar{\mathbf{x}}) - n\phi_l)}}{q^n(\mathbf{x}, \mathbf{y})^s e^{\sum_l \bar{r}_l (a_l^n(\mathbf{x}) - n\phi_l)}} \right)^\rho \quad (62)$$

for any real numbers  $\{r_l\}$  and  $\{\bar{r}_l\}$ . Expanding each term in the outer summation as product from  $i = 1$  to  $n$ , we obtain

$$\bar{p}_e(n, M) \leq \frac{e^{\rho \sum_l (|r_l| + |\bar{r}_l|)\delta}}{\mu_n^{1+\rho}} M^\rho \left( \sum_{\mathbf{x}, \mathbf{y}} Q(\mathbf{x}) W(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}}} Q(\bar{\mathbf{x}}) q(\bar{\mathbf{x}}, \mathbf{y})^s e^{\sum_l r_l (a_l(\bar{\mathbf{x}}) - \phi_l)}}{q(\mathbf{x}, \mathbf{y})^s e^{\sum_l \bar{r}_l (a_l(\mathbf{x}) - \phi_l)}} \right)^\rho \right)^n. \quad (63)$$

Since  $\mu_n$  decays to zero subexponentially in  $n$  (cf. Proposition 1) and  $e^{\rho \sum_l (|r_l| + |\bar{r}_l|)\delta}$  is independent of  $n$ , we conclude that the factor

$$\frac{e^{\rho \sum_l (|r_l| + |\bar{r}_l|)\delta}}{\mu_n^{1+\rho}} \quad (64)$$

does not affect the exponential behavior of (63). We therefore recover the error exponent in (53).

Unlike the proof of Theorem 4, this direct derivation does not prove ensemble tightness. A similar statement holds for the above-mentioned direct derivations of (51)–(52). However, these have the advantage of extending to non-finite alphabets. Under constant-composition random coding, the input distribution  $Q$  must have finite support, but the output alphabet may be infinite. For the cost-constrained ensemble, we only require that the second moments of the cost functions are finite, in accordance with Proposition 1.

### A. Connections Between the Error Exponents

The constraints on  $P_{XY}$  and  $\tilde{P}_{XY}$  in (51)–(53) are given by the sets defined in (42)–(47). Since the constraint  $\mathbb{E}_P[a_l(X)] = \phi_l$  holds by definition when  $P_X = Q$ , we have that  $\mathcal{S}^{\text{iid}} \subseteq \mathcal{S}^{\text{cost}}(\{a_l\}) \subseteq \mathcal{S}^{\text{cc}}(Q)$  for any given set of cost functions  $\{a_l(\cdot)\}$  and input distribution  $Q$ . A similar observation applies to the constraints on  $\tilde{P}_X$ , and it follows that

$$E_r^{\text{iid}}(Q, R) \leq E_r^{\text{cost}}(Q, R, \{a_l\}) \leq E_r^{\text{cc}}(Q, R). \quad (65)$$

This indicates that the constant-composition ensemble yields the best error exponent of the three ensembles under consideration.

By choosing  $L = 1$  and setting  $r_1 = \bar{r}_1 = 1$  in (56), we obtain the inequality

$$E_r^{\text{cost}'}(Q, R, a_1) \leq E_r^{\text{cost}}(Q, R, a_1). \quad (66)$$

Furthermore, for any set of cost functions  $\{a_l\}_{l=1}^L$ , we have

$$E_r^{\text{cost}}(Q, R, \{a_l\}_{l=1}^{L-1}) \leq E_r^{\text{cost}}(Q, R, \{a_l\}_{l=1}^L). \quad (67)$$

That is, additional cost functions can never decrease the error exponent.

In the case that  $a_1(x)$  does not depend on  $x$ , we obtain

$$E_r^{\text{cost}}(Q, R, a_1) = E_r^{\text{cost}'}(Q, R, a_1) = E_r^{\text{iid}}(Q, R), \quad (68)$$

and hence we have in general that

$$E_r^{\text{iid}}(Q, R) \leq \sup_{a_1(\cdot)} E_r^{\text{cost}'}(Q, R, a_1). \quad (69)$$

The following theorem gives the connection between  $E_r^{\text{cost}'}$  and  $E_r^{\text{cc}}$ , and shows that the two are equal after the optimization over  $Q$  and  $a_1(\cdot)$ . This result is analogous to a connection between the i.i.d. ensemble and constant-composition ensemble under ML decoding [24, Theorem 3.4]. The general steps of the proof are similar to that of [24], but the details are more involved.

**Theorem 5.** *The function  $E_0^{\text{cc}}(Q, \rho)$  can be expressed as*

$$E_0^{\text{cc}}(Q, \rho) = \max_{\tilde{Q}} \sup_{a_1(\cdot)} E_0^{\text{cost}'}(\tilde{Q}, \rho, a_1) - (1 + \rho)D(Q\|\tilde{Q}). \quad (70)$$

Consequently,

$$\max_Q E_0^{\text{cc}}(Q, \rho) = \max_Q \sup_{a_1(\cdot)} E_0^{\text{cost}'}(Q, \rho, a_1). \quad (71)$$

*Proof:* The proof of (70) is given in Appendix C. We obtain (71) from (70) since  $D(Q\|\tilde{Q}) \geq 0$  with equality if and only if  $Q = \tilde{Q}$ . ■

It follows from (65) and Theorem 5 that  $E_r^{\text{cost}'}$  is tight with respect to the ensemble average when the optimal values of both  $Q$  and  $a_1(\cdot)$  are used. Hence, although  $E_r^{\text{cost}}$  is a tighter error exponent than  $E_r^{\text{cost}'}$  in general, it does not improve on the best achievable error exponent using cost-constrained random coding. Furthermore, both exponents can be used to prove the achievability of the LM rate and no better. However, the refined exponent

$E_r^{\text{cost}}$  is useful in the case that one does not have complete freedom in choosing  $Q$  and  $a(\cdot)$ , or when the exact optimization over each is not feasible. For example, if the codebook designer does not know the channel, then the objective in (58) cannot be computed in order to perform the optimization. Furthermore, even if the channel and metric are known, some input distributions may be ruled out due to system constraints (e.g. power limitations).

The following theorem gives two further connections between the error exponents in the case of ML decoding.

**Theorem 6.** *Under the ML metric  $q(x, y) = W(y|x)$ , we have*

$$\sup_{a(\cdot)} E_r^{\text{cost}'}(Q, R, a) = E_r^{\text{iid}}(Q, R) \quad (72)$$

$$\sup_{a(\cdot)} E_r^{\text{cost}}(Q, R, a) = E_r^{\text{cc}}(Q, R). \quad (73)$$

*Proof:* We obtain (72) by optimizing the objective in (54) over  $s$ , and optimizing the objective in (58) over  $s$  and  $a(\cdot)$ . It was shown in [1, Ex. 5.6] that the optimal value of  $s$  in (54) is equal to  $\frac{1}{1+\rho}$ . Following the same steps, we obtain that the optimal value of  $s$  in (58) is also equal to  $\frac{1}{1+\rho}$ , and the optimal cost function  $a(x)$  does not depend on  $x$ . Combining these results, we obtain (72).

From (65), it suffices to prove that  $\sup_{a(\cdot)} E_r^{\text{cost}} \geq E_r^{\text{cc}}$  in order to prove (73). To this end, we will show that  $\sup_{a(\cdot)} E_0^{\text{cost}}(Q, \rho, a) \geq E_0^{\text{cc}}(Q, \rho)$  for all  $\rho \in [0, 1]$ . The case  $\rho = 0$  is trivial, so we assume that  $\rho > 0$ . We set  $L = 1$ ,  $\bar{r}_1 = 1$  and  $r_1 = \frac{-1}{\rho}$  and write the expectation inside the logarithm of (56) as

$$\sum_{x,y} Q(x)W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{\phi_a}} \right)^\rho \frac{e^{a(x)}}{e^{\phi_a}}. \quad (74)$$

Introducing the distribution  $\tilde{Q}(x) = \frac{Q(x)e^{a(x)}}{\mathbb{E}_Q[e^{a(X)}]}$ , we can write (74) as

$$\left( \frac{\mathbb{E}_Q[e^{a(X)}]}{e^{\phi_a}} \right)^{1+\rho} \sum_{x,y} \tilde{Q}(x)W(y|x) \left( \frac{\sum_{\bar{x}} \tilde{Q}(\bar{x})q(\bar{x}, y)^s}{q(x, y)^s} \right)^\rho. \quad (75)$$

The summation in (75) coincides with the expectation inside the logarithm of (54). Furthermore, we have

$$D(Q\|\tilde{Q}) = \sum_x Q(x) \log \frac{\mathbb{E}_Q[e^{a(X)}]}{e^{a(x)}} \quad (76)$$

$$= \log \mathbb{E}_Q[e^{a(X)}] - \sum_x Q(x)a(x) \quad (77)$$

$$= \log \frac{\mathbb{E}_Q[e^{a(X)}]}{e^{\phi_a}}, \quad (78)$$

where (76) follows from the definition of  $\tilde{Q}$ , and (78) follows from the definition of  $\phi_a$ . Substituting (75) and (78) into (56) and noting that a suitable choice of  $a(\cdot)$  can yield any distribution  $\tilde{Q}$  with the same support as  $Q$ , we obtain

$$\sup_{a(\cdot)} E_0^{\text{cost}}(Q, \rho, a) \geq \max_{\tilde{Q}} E_0^{\text{iid}}(\tilde{Q}, \rho) - (1 + \rho)D(Q\|\tilde{Q}). \quad (79)$$

Taking account of (72) and Theorem 5, the right-hand side of (79) is equal to  $E_0^{\text{cc}}(Q, \rho)$ , and the proof is complete. ■



Under ML decoding, the exponent  $E_r^{\text{iid}}$  is simply Gallager's random-coding error exponent [1, Sec. 5.6], while  $E_r^{\text{cc}}$  is Csisz  r's random-coding error exponent [14, Sec. 10] (see also [24] for an equivalent dual expression). The two exponents are equal after the optimization over  $Q$ , but Csisz  r's exponent can be higher for a given  $Q$  [15]. Theorem 6 shows that using an optimized cost function  $a(\cdot)$ , we can achieve Csisz  r's exponent in the matched setting using one suitably chosen cost function, rather than constant-composition codes. The following theorem shows that a similar result holds in the mismatched setting using two suitably chosen cost functions.

**Theorem 7.** *For any decoding metric  $q(x, y)$ , we have*

$$\sup_{a_1(\cdot), a_2(\cdot)} E_r^{\text{cost}}(Q, R, \{a_1, a_2\}) = E_r^{\text{cc}}(Q, R). \quad (80)$$

*Proof:* Setting  $L = 2$  and choosing  $r_1 = \bar{r}_1 = \bar{r}_2 = 1$  and  $r = \frac{-1}{\rho}$ , the expectation in (56) becomes

$$\sum_{x, y} Q(x)W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a_1(\bar{x})+a_2(\bar{x})}}{q(x, y)^s e^{a_1(x)+\phi_2}} \right)^\rho \frac{e^{a_2(x)}}{e^{\phi_2}}. \quad (81)$$

Defining  $\tilde{Q}(x) = \frac{Q(x)e^{a_2(x)}}{\mathbb{E}_Q[e^{a_2(X)}]}$  and following identical steps to (75)–(79), we obtain

$$\sup_{a_1(\cdot), a_2(\cdot)} E_0^{\text{cost}}(Q, \rho, \{a_1, a_2\}) \geq \max_Q \sup_{a_1(\cdot)} E_0^{\text{cost}'}(\tilde{Q}, \rho, a_1) - (1 + \rho)D(Q\|\tilde{Q}). \quad (82)$$

By Theorem 5, the right-hand side of (82) is equal to  $E_0^{\text{cc}}(Q, \rho)$ , and we have thus recovered the exponent  $E_r^{\text{cc}}$ . The matching lower bound on  $E_r^{\text{cost}}$  is given in (65). ■

We conclude this subsection by showing that (80) can be derived directly by a suitable choice of the cost functions  $a_1(\cdot)$  and  $a_2(\cdot)$ . An advantage of this approach is that it provides an explicit formula for  $a_2(\cdot)$  in terms of  $s$  and  $a_1(\cdot)$ . Furthermore, from the proof of Theorem 6, we can set  $s = \frac{1}{1+\rho}$  and  $a_1(x) = 0$  in the case of ML decoding without loss of optimality, and hence all of the optimization parameters are known analytically in this case.

Fix the input distribution  $Q$ , and the parameters  $\rho, s \geq 0$  and  $a(x)$ . We set  $a_1(x) = a(x)$  and

$$a_2(x) = \log \sum_y W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right)^\rho. \quad (83)$$

Using (61), we have

$$\bar{p}_e \leq \frac{e^{2\rho\delta}}{\mu_n^\rho} \sum_{\mathbf{x} \in \mathcal{D}_n, \mathbf{y}} Q(\mathbf{x})W^n(\mathbf{y}|\mathbf{x}) \left( M \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}})q^n(\bar{\mathbf{x}}, \mathbf{y})^s e^{a^n(\bar{\mathbf{x}})}}{q^n(\mathbf{x}, \mathbf{y})^s e^{a^n(\mathbf{x})}} \right)^\rho \quad (84)$$

$$\leq \frac{e^{2\rho\delta}}{\mu_n^\rho} M^\rho \max_{\mathbf{x} \in \mathcal{D}_n} \sum_{\mathbf{y}} W^n(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}})q^n(\bar{\mathbf{x}}, \mathbf{y})^s e^{a^n(\bar{\mathbf{x}})}}{q^n(\mathbf{x}, \mathbf{y})^s e^{a^n(\mathbf{x})}} \right)^\rho, \quad (85)$$

where  $Q^n(\mathbf{x}) \triangleq \prod_{i=1}^n Q(x_i)$  and  $a^n(\mathbf{x}) \triangleq \sum_{i=1}^n a(x_i)$ . We write the summation in (85) as

$$\exp \left( \log \sum_{\mathbf{y}} W^n(\mathbf{y}|\mathbf{x}) \left( \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}})q^n(\bar{\mathbf{x}}, \mathbf{y})^s e^{a^n(\bar{\mathbf{x}})}}{q^n(\mathbf{x}, \mathbf{y})^s e^{a^n(\mathbf{x})}} \right)^\rho \right). \quad (86)$$

The constraint on  $a_2(x)$  in the definition of  $\mathcal{D}_n$  in (16) implies that, for all  $\mathbf{x} \in \mathcal{D}_n$ , the logarithm in (86) differs from its mean by no more than  $\delta$ . Expanding each term in the logarithm as a product from 1 to  $n$ , we see that this mean is simply  $nE_0^{\text{cc}}$  (see (55)), and thus (86) is upper bounded by

$$\exp(nE_0^{\text{cc}}(Q, \rho) + \delta). \quad (87)$$

The derivation is concluded by substituting (87) into (85), using Proposition 1, and maximizing over  $\rho \in [0, 1]$ ,  $s \geq 0$  and  $a(\cdot)$ .

This analysis in (83)–(87) extends immediately to general alphabets, provided that the second moments of  $a_1(X)$  and  $a_2(X)$  are finite under  $X \sim Q$  (see Proposition 1). In the discrete memoryless setting, we relied on the primal expressions to show that we can do no better than  $E_r^{cc}$ . To see that this is true more generally, we lower bound (55) as follows:

$$E_0^{cc}(Q, \rho) \geq \mathbb{E} \left[ -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{\sum_{l=1}^L \bar{r}_l(a_l(\bar{X}) - \phi_l)} | Y]}{q(X, Y)^s e^{\sum_{l=1}^L \bar{r}_l(a_l(X) - \phi_l)}} \right)^\rho \middle| X \right] \right] \quad (88)$$

$$= \mathbb{E} \left[ -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{\sum_{l=1}^L \bar{r}_l(a_l(\bar{X}) - \phi_l)} | Y]}{q(X, Y)^s e^{\sum_{l=1}^L r_l(a_l(X) - \phi_l)}} \right)^\rho \middle| X \right] \right] \quad (89)$$

$$\geq -\log \mathbb{E} \left[ \left( \frac{\mathbb{E}[q(\bar{X}, Y)^s e^{\sum_{l=1}^L \bar{r}_l(a_l(\bar{X}) - \phi_l)} | Y]}{q(X, Y)^s e^{\sum_{l=1}^L r_l(a_l(X) - \phi_l)}} \right)^\rho \right], \quad (90)$$

where (88) follows for any  $s > 0$ ,  $\{a_l\}$ , and  $\{\bar{r}_l\}$  by lower bounding the supremum over  $a(x)$  by the particular choice  $\sum_{l=1}^L \bar{r}_l(a_l(X_l) - \phi_l)$  with  $\phi_l = \mathbb{E}_Q[a_l(X)]$ , (89) follows for any  $\{r_l\}$  by expanding the logarithm of products as a sum of logarithms and noting that the term  $e^{\sum_{l=1}^L r_l(a_l(X) - \phi_l)}$  has no effect on the resulting expression regardless of what values  $\{r_l\}$  take, and (90) follows from Jensen's inequality. By optimizing (90) over  $s$ ,  $\{a_l\}$ ,  $\{r_l\}$  and  $\{\bar{r}_l\}$ , we obtain the desired result. In summary, we have shown that  $L = 2$  cost functions suffice to achieve  $E_r^{cc}$ , and we can do no better by using more cost functions.

### B. Numerical Results

In this section, we plot the exponents for the channel defined in (40) under both the minimum Hamming distance and ML decoding metrics, again using the parameters  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$  and  $\delta_2 = 0.25$ . We set  $Q = (0.1, 0.3, 0.6)$ , which we have intentionally chosen suboptimally to highlight the differences between the error exponents when the input distribution is fixed. Under these parameters we have  $I^{\text{GMI}}(Q) = 0.387$ ,  $I^{\text{LM}}(Q) = 0.449$  and  $I(X; Y) = 0.471$  bits/use.

We evaluate the exponents using the optimization software YALMIP [25]. We allow each of the optimization parameters (e.g.  $s$  and  $a(\cdot)$ ) to vary with  $R$ . The exponent  $E_r^{\text{cost}'}$  is optimized over  $a_1(\cdot)$ , performing the optimization jointly with  $s$  in (58). The cost function for  $E_r^{\text{cost}}$  ( $L = 1$ ) is chosen using an alternating optimization between  $a_1(\cdot)$  and  $(s, r, \bar{r})$  in (56), using the  $a_1(\cdot)$  which maximizes  $E_0^{\text{cost}'}$  as a starting point and terminating when the change in the exponent between iterations becomes negligible. To demonstrate the effectiveness of this procedure, the cost functions for  $E_r^{\text{cost}}$  ( $L = 2$ ) are chosen using an alternating optimization between  $(a_1, a_2)$  and  $(s, r_1, r_2, \bar{r}_1, \bar{r}_2)$ , initially setting both  $a_1(\cdot)$  and  $a_2(\cdot)$  to be equal to the  $a_1(\cdot)$  which maximizes  $E_0^{\text{cost}'}$ . From Theorem 7, we expect this to recover the exponent  $E_r^{cc}$ .

From Figure 3, we see that  $E_r^{cc}$  and  $E_r^{\text{cost}}$  ( $L = 2$ ) are indistinguishable, indicating that the alternating optimization technique was effective in finding the true exponent. The exponent  $E_r^{\text{cost}}$  ( $L = 1$ ) is only marginally lower, while the gap to  $E_r^{\text{cost}'}$  is larger. The exponent  $E_r^{\text{iid}}$  is not only lower than each of the other exponents, but also yields

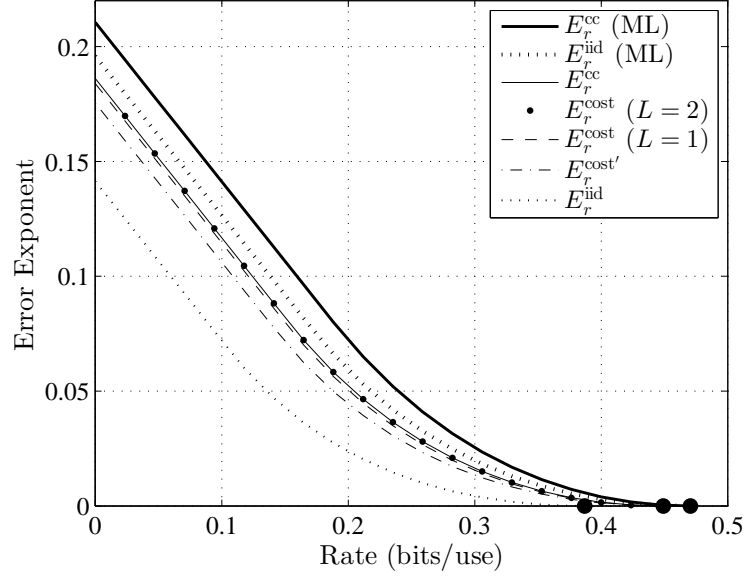


Figure 3. Error exponents for the channel defined in (40) with  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$  and  $Q = (0.1, 0.3, 0.6)$ . The mismatched decoder uses the minimum Hamming distance metric given in (41). The corresponding achievable rates  $I^{\text{GMI}}(Q)$ ,  $I^{\text{LM}}(Q)$  and  $I(X; Y)$  are respectively marked on the horizontal axis.

a worse achievable rate. This example demonstrates that for a fixed  $Q$ , the refined exponent  $E_r^{\text{cost}}(L = 1)$  can outperform  $E_r^{\text{cost}'}$  even when  $a_1(\cdot)$  is optimized.

## V. SECOND-ORDER CODING RATES

In the matched setting, the finite-length performance limits of a channel are characterized by  $M^*(n, \epsilon)$ , defined to be the maximum number of codewords of length  $n$  yielding an error probability not exceeding  $\epsilon$  for some encoder-decoder pair. For channels satisfying the strong converse,  $M^*(n, \epsilon)$  satisfies

$$\log M^*(n, \epsilon) = nC + o(n) \quad (91)$$

for all  $\epsilon \in (0, 1)$ , where  $C$  is the channel capacity. The problem of finding higher order terms in this expansion was studied by Strassen [26], and later revisited by Polyanskiy *et al.* [18] and Hayashi [27], among others. For DMCs, the second-order expansion is of the form

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + O(\log n), \quad (92)$$

where  $V$  is a constant known as the channel dispersion. The expansion in (92) is often referred to as the normal approximation, and is said to describe the second-order coding rate.

In this section, we present achievable second-order coding rates for the ensembles given in Section I, i.e. expansions of the form (92) with the equality replaced by  $\geq$ . To distinguish between the ensembles, we define  $M^{\text{iid}}(Q, n, \epsilon)$ ,  $M^{\text{cc}}(Q, n, \epsilon)$  and  $M^{\text{cost}}(Q, n, \epsilon)$  to be the maximum number of codewords of length  $n$  such that

the random-coding error probability does not exceed  $\epsilon$  for the i.i.d. ensemble, constant-composition ensemble and cost-constrained ensemble respectively, using the input distribution  $Q$ . For the cost-constrained ensemble, we choose  $L = 2$  and allow the cost functions to be optimized. More specifically, we will fix one cost function arbitrarily and choose the other one in terms of  $W$ ,  $q$ ,  $Q$ , and the first cost function.

A key quantity in the second-order analysis for ML decoding is the information density, given by [28]

$$i(x, y) \triangleq \log \frac{W(y|x)}{\sum_x Q(x)W(y|x)}, \quad (93)$$

where  $Q$  is a given input distribution. In the mismatched setting, two generalizations of  $i(x, y)$  are given by

$$i_s(x, y) \triangleq \log \frac{q(x, y)^s}{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s} \quad (94)$$

$$i_{s,a}(x, y) \triangleq \log \frac{q(x, y)^s e^{a(x)}}{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}, \quad (95)$$

where  $s \geq 0$  and  $a(\cdot)$  are parameters. Similarly to Section IV, we will see that  $a(\cdot)$  represents a mathematical optimization parameter for the constant-composition ensemble, whereas for the cost-constrained ensemble it represents one of the cost functions. The information density is recovered from (94)–(95) by setting  $q(x, y) = W(y|x)$ ,  $s = 1$ , and  $a(x) = 0$ . We write

$$i_s^n(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n i_s(x_i, y_i) \quad (96)$$

$$i_{s,a}^n(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n i_{s,a}(x_i, y_i) \quad (97)$$

as well as the usual notations  $Q^n(\mathbf{x}) \triangleq \prod_{i=1}^n Q(x_i)$  and  $a^n(\mathbf{x}) \triangleq \sum_{i=1}^n a(x_i)$ . We define the quantities

$$I_s(Q) \triangleq \mathbb{E}[i_s(X, Y)] \quad (98)$$

$$U_s(Q) \triangleq \text{Var}[i_s(X, Y)] \quad (99)$$

$$I_{s,a}(Q) \triangleq \mathbb{E}[i_{s,a}(X, Y)] \quad (100)$$

$$V_{s,a}(Q) \triangleq \mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]], \quad (101)$$

where  $(X, Y) \sim Q \times W$ . From (4) and (9), we see that the GMI (respectively, LM rate) is equal to  $I_s(Q)$  (respectively,  $I_{s,a}(Q)$ ) after optimizing over  $s$  (respectively,  $s$  and  $a(\cdot)$ ).

We state the second-order coding rates for the i.i.d. and constant-composition ensembles without proof, since their analysis is similar to that of the matched setting [18], [26], [29]. For the i.i.d. ensemble we obtain

$$\log M^{\text{iid}}(Q, n, \epsilon) \geq nI_s(Q) - \sqrt{nU_s(Q)} Q^{-1}(\epsilon) + O(1) \quad (102)$$

for any  $Q$  and  $s \geq 0$ . Similarly, for the constant-composition ensemble, we obtain

$$\log M^{\text{cc}}(Q, n, \epsilon) \geq nI_{s,a}(Q) - \sqrt{nV_{s,a}(Q)} Q^{-1}(\epsilon) + O(\log n) \quad (103)$$

for any  $Q$ ,  $s \geq 0$ , and  $a(\cdot)$ . These expressions can also be obtained by following the proof of Theorem 8 below. The corresponding results for the matched setting are obtained by setting  $s = 1$  and  $a(x) = 0$ .

The constant-composition ensemble yields two advantages over that of the i.i.d. ensemble: the additional parameter  $a(\cdot)$ , and a conditional (rather than unconditional) variance in the second-order term. In the matched setting, these advantages vanish, since the choice  $a(x) = 0$  is optimal, and the conditional and unconditional variances coincide under the capacity-achieving input distribution [18]. Nevertheless, when system cost constraints are present, the conditional variance can be strictly higher than the unconditional variance after optimizing  $Q$ , even under ML decoding.

The main result of this subsection is the following theorem, which states that the cost-constrained ensemble yields the same second-order expansion as (103). We first present a proof for DMCs which yields a  $o(\sqrt{n})$  third-order term, and then discuss the changes required to obtain a  $O(\log n)$  third-order term and handle general alphabets. Our proof differs from the usual proof using threshold-based random-coding bounds [18], [26], but the latter approach can also be used in the present setting [30]. Our analysis can be interpreted as performing a normal approximation of the bound  $\text{rcu}_s$  in (26).

**Theorem 8.** *Using cost-constrained random-coding with  $L = 2$ ,*

$$\log M^{\text{cost}}(Q, n, \epsilon) \geq nI_{s,a}(Q) - \sqrt{nV_{s,a}(Q)} Q^{-1}(\epsilon) + o(\sqrt{n}) \quad (104)$$

for any  $s \geq 0$  and  $a(\cdot)$ .

*Proof:* Throughout the proof, we make use of the random variables  $(X, Y, \bar{X}) \sim Q(x)W(y|x)Q(\bar{x})$  and  $(\mathbf{X}, \mathbf{Y}, \bar{\mathbf{X}}) \sim P_{\mathbf{X}}(\mathbf{x})W^n(\mathbf{y}|\mathbf{x})P_{\mathbf{X}}(\bar{\mathbf{x}})$ . Probabilities and expectations containing a realization  $x$  of  $X$  are implicitly defined to be conditioned on the event  $X = x$ , and similarly for realizations  $\mathbf{x}$  of  $\mathbf{X}$ .

We choose the cost functions

$$a_1(x) = a(x) \quad (105)$$

$$a_2(x) = \mathbb{E}[i_{s,a}(x, Y)]. \quad (106)$$

We weaken  $\text{rcu}_s$  in (26) as follows:

$$\text{rcu}_s(n, M) = \mathbb{E} \left[ \min \left\{ 1, (M-1) \frac{\sum_{\bar{\mathbf{x}}} P_{\mathbf{X}}(\bar{\mathbf{x}}) q^n(\bar{\mathbf{X}}, \mathbf{Y})^s}{q^n(\mathbf{X}, \mathbf{Y})^s} \right\} \right] \quad (107)$$

$$\leq \mathbb{E} \left[ \min \left\{ 1, M e^{2\delta} \frac{\sum_{\bar{\mathbf{x}}} P_{\mathbf{X}}(\bar{\mathbf{x}}) q^n(\bar{\mathbf{X}}, \mathbf{Y})^s e^{a^n(\bar{\mathbf{X}})}}{q^n(\mathbf{X}, \mathbf{Y})^s e^{a^n(\mathbf{X})}} \right\} \right] \quad (108)$$

$$\leq \mathbb{E} \left[ \min \left\{ 1, \frac{M e^{2\delta}}{\mu_n} \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}}) q^n(\bar{\mathbf{X}}, \mathbf{Y})^s e^{a^n(\bar{\mathbf{X}})}}{q^n(\mathbf{X}, \mathbf{Y})^s e^{a^n(\mathbf{X})}} \right\} \right] \quad (109)$$

$$= \mathbb{P} \left[ i_{s,a}^n(\mathbf{X}, \mathbf{Y}) + \log U \leq \log \frac{M e^{2\delta}}{\mu_n} \right], \quad (110)$$

where (108) follows from the constraints on  $a^n(\mathbf{x})$  in (16), (109) follows by substituting the random-coding distribution in (15) and summing over all  $\bar{\mathbf{x}}$  instead of just  $\bar{\mathbf{x}} \in \mathcal{D}_n$ , and (110) follows from the definition of  $i_{s,a}$  and the identity  $\mathbb{E}[\min\{1, A\}] = \mathbb{P}[A > U]$ , where  $U$  is uniform on  $(0, 1)$  and independent of  $A$ .

For any set  $\mathcal{A}_n \subseteq \mathcal{D}_n$ , (110) implies

$$\bar{p}_e(n, M) \leq \mathbb{P}[\mathbf{X} \notin \mathcal{A}_n] + \max_{\mathbf{x} \in \mathcal{A}_n} \mathbb{P}\left[i_{s,a}^n(\mathbf{x}, \mathbf{Y}) + \log U \leq \log \frac{Me^{2\delta}}{\mu_n}\right]. \quad (111)$$

Choosing  $M$  such that

$$\log M = nI_{s,a}(Q) - \log \mu_n - 2\delta - \beta \quad (112)$$

for some  $\beta$ , we obtain

$$\bar{p}_e(n, M) \leq \mathbb{P}[\mathbf{X} \notin \mathcal{A}_n] + \max_{\mathbf{x} \in \mathcal{A}_n} \mathbb{P}\left[i_{s,a}^n(\mathbf{x}, \mathbf{Y}) + \log U \leq nI_{s,a}(Q) - \beta\right]. \quad (113)$$

Choosing

$$\mathcal{A}_n \triangleq \left\{ \mathbf{x} \in \mathcal{D}_n : \text{Var}[i_{s,a}^n(\mathbf{x}, \mathbf{Y})] \leq n\left(\mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]] + \delta'\right) \right\} \quad (114)$$

for some  $\delta' > 0$ , we obtain

$$\mathbb{P}[\mathbf{X} \notin \mathcal{A}_n] = \sum_{\mathbf{x} \in \mathcal{D}_n} P_{\mathbf{X}}(\mathbf{x}) \mathbb{1}\left\{ \text{Var}[i_{s,a}^n(\mathbf{x}, \mathbf{Y})] > n\left(\mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]] + \delta'\right) \right\} \quad (115)$$

$$\leq \frac{1}{\mu_n} \sum_{\mathbf{x}} Q^n(\mathbf{x}) \mathbb{1}\left\{ \text{Var}[i_{s,a}^n(\mathbf{x}, \mathbf{Y})] > n\left(\mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]] + \delta'\right) \right\} \quad (116)$$

$$\leq e^{-n(E(\delta') + o(1))}, \quad (117)$$

where (117) holds for some  $E(\delta') > 0$  by Hoeffding's inequality for bounded random variables [31], and since  $\mu_n$  decays at most polynomially in  $n$ . Furthermore, from the definitions of  $\mathcal{A}_n$  and  $\mathcal{D}_n$  and the choice of the cost function in (106), we have for any  $\mathbf{x} \in \mathcal{A}_n$  that

$$\mathbb{E}[i_{s,a}^n(\mathbf{x}, \mathbf{Y})] \geq n\mathbb{E}[i_{s,a}(X, Y)] - \delta \quad (118)$$

$$\text{Var}[i_{s,a}^n(\mathbf{x}, \mathbf{Y})] \leq n\left(\mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]] + \delta'\right), \quad (119)$$

where  $\delta$  is the parameter to the ensemble appearing in (16). Since the moments of  $\log U$  are finite, this implies

$$\mathbb{E}[i_{s,a}^n(\mathbf{x}, \mathbf{Y}) + \log U] \geq n\mathbb{E}[i_{s,a}(X, Y)] + O(1) \quad (120)$$

$$\text{Var}[i_{s,a}^n(\mathbf{x}, \mathbf{Y}) + \log U] \leq n\left(\mathbb{E}[\text{Var}[i_{s,a}(X, Y) | X]] + \delta'\right) + O(1). \quad (121)$$

Thus, applying (117) and the Berry-Esseen theorem for independent and non-identically distributed random variables [16, Sec. XVI.5] to (113), we obtain

$$\bar{p}_e(n, M) \leq \mathbb{Q}\left(\frac{\beta + O(1)}{\sqrt{n(V_{s,a}(Q) + \delta') + O(1)}}\right) + O\left(\frac{1}{\sqrt{n}}\right), \quad (122)$$

where we have used the fact that the third absolute moment of  $i_{s,a}(X, Y)$  is always bounded when the alphabets  $\mathcal{X}$  and  $\mathcal{Y}$  are finite [18, Lemma 46]. By straightforward rearrangements and a first-order Taylor expansion of the square root function and the  $\mathbb{Q}^{-1}$  function, we obtain

$$\beta \leq \sqrt{n(V_{s,a}(Q) + \delta')} \mathbb{Q}^{-1}(\bar{p}_e) + O(1) \quad (123)$$

$$= \sqrt{nV_{s,a}(Q)} \mathbb{Q}^{-1}(\bar{p}_e) + O(\sqrt{n}\delta'). \quad (124)$$

The proof is concluded by combining (112) and (124), taking  $\delta' \rightarrow 0$ , and noting from Proposition 1 that  $\log \mu_n = O(\log n)$ .  $\blacksquare$

The third-order term  $o(\sqrt{n})$  in (104) can be improved to  $O(\log n)$  by introducing a third cost function  $a_3(x) = \text{Var}[i_{s,a}(x, Y)]$ , thus constraining  $\text{Var}[i_{s,a}^n(x, \mathbf{Y})]$  to be close to its mean, similarly to  $\mathbb{E}[i_{s,a}^n(x, \mathbf{Y})]$  in (118). We believe that, after optimizing over  $s$  and  $a(\cdot)$ , the expansions in (102)–(104) hold with equality, and thus the analysis is tight with respect to the ensemble average. However, thus far we do not have a proof which holds for an arbitrary metric  $q(x, y)$ .

The expansion in (104) applies to channels with general alphabets, under some technical assumptions. In order to apply Proposition 1, we require each of the cost functions to have a finite second moment. Furthermore, Hoeffding's inequality in (117) cannot be applied to unbounded random variables. Defining  $v_{s,a}(x) \triangleq \text{Var}[i_{s,a}(x, Y)]$  and  $\mathbf{X}' \sim Q^n$ , it must be checked that the quantity

$$\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n v_{s,a}(X'_i) - V_{s,a}(Q) \geq \delta'\right] \quad (125)$$

decays to zero sufficiently quickly (not necessarily exponentially fast) for all  $\delta' \geq 0$ . Similarly, when  $i_{s,a}(x, y)$  is not bounded, we must include a constraint on the third absolute moment  $t_{s,a}(x) \triangleq \mathbb{E}[|i_{s,a}(x, Y) - I_{s,a}(Q)|^3]$  in the set  $\mathcal{A}_n$ , and ensure that the analogous quantity to (125) decays to zero sufficiently fast for some  $\delta' > 0$ . Alternatively, one can take the approach outlined in the previous paragraph and let  $v_{s,a}(\cdot)$  and  $t_{s,a}(\cdot)$  be cost functions in the ensemble. These will have a finite second moment provided that the sixth moment of  $i_{s,a}(X, Y)$  is finite, and in this case no further assumptions are required.

## VI. SADDLEPOINT APPROXIMATIONS

A limitation of the random-coding bounds given in Section III is that they are often too complex to compute. In this section, we show that saddlepoint approximations [32] can be used to accurately characterize certain random-coding bounds with a computational complexity comparable to that of the error exponents and second-order rates.

### A. Motivation

Chernoff-type bounds provide an estimate of the tail probability of a random variable via the cumulant transform, namely  $\mathbb{P}[\zeta > z] \approx e^{\kappa(\hat{\tau}) - \hat{\tau}z}$ , where  $\kappa(\tau) = \log \mathbb{E}[e^{\tau\zeta}]$  and  $\hat{\tau} = \arg \min_{\tau} \kappa(\tau) - \tau z$ . Saddlepoint approximations provide more accurate estimates of the form  $\mathbb{P}[\zeta > z] \approx \alpha(\kappa, \hat{\tau}) \cdot e^{\kappa(\hat{\tau}) - \hat{\tau}z}$ . In its classical form, the coefficient  $\alpha(\kappa, \hat{\tau})$  is given by [32, Sec. 2.2]

$$\alpha(\kappa, \hat{\tau}) = \frac{1}{\hat{\tau} \sqrt{2\pi\kappa''(\hat{\tau})}}. \quad (126)$$

This expression has been used previously in the information theory literature, e.g. see [1, Appendix 5A]. It often yields accurate approximations, but it is inaccurate when  $\hat{\tau}$  is close to zero, which occurs when  $z$  is close to the mean  $\mathbb{E}[\zeta]$ . In such cases, and more generally, a better coefficient is given by

$$\alpha(\kappa, \hat{\tau}) = \frac{1}{2} \text{erfcx}\left(\hat{\tau} \sqrt{\frac{\kappa''(\hat{\tau})}{2}}\right), \quad (127)$$

where  $\operatorname{erfcx}(x) \triangleq \operatorname{erfc}(x) \exp(x^2)$ . An asymptotic expansion of  $\operatorname{erfc}(x)$  as  $x \rightarrow \infty$  recovers (126). However, as  $z$  approaches  $\mathbb{E}[\zeta]$  from above, we have  $\hat{\tau} \rightarrow \frac{\mathbb{E}[\zeta] - z}{\kappa''(0)}$ , where  $\kappa''(0)$  is the variance of  $\zeta$ . In this case, we have

$$\mathbb{P}[\zeta > z] \approx \frac{1}{2} \operatorname{erfcx}\left(\hat{\tau} \sqrt{\frac{\kappa''(\hat{\tau})}{2}}\right) e^{\kappa(\hat{\tau}) - \hat{\tau}z} \quad (128)$$

$$\approx Q\left(\frac{\mathbb{E}[\zeta] - z}{\sqrt{\kappa''(0)}}\right). \quad (129)$$

We thus recover a normal approximation to the probability for values of  $z$  close to the mean  $\mathbb{E}[\zeta]$ .

### B. Approximation for the i.i.d. Ensemble

In this subsection, we consider the approximation of the weakened RCU bound  $\operatorname{rcu}_s(n, M)$  under the i.i.d. ensemble. We begin by presenting a heuristic derivation of the saddlepoint approximation. A rigorous derivation including error bounds is given in the proof of Theorem 9 below.

Using the identity  $\mathbb{E}[\min\{1, A\}] = \mathbb{P}[A > U]$ , where  $U$  is uniform on  $(0, 1)$  and independent of  $A$ , we can express  $\operatorname{rcu}_s(n, M)$  as

$$\operatorname{rcu}_s(n, M) = \mathbb{P}[\zeta_n \geq 0], \quad (130)$$

where

$$\zeta_n \triangleq \log \frac{M-1}{U} - i_s^n(\mathbf{X}, \mathbf{Y}), \quad (131)$$

and  $i_s^n(\mathbf{X}, \mathbf{Y})$  is defined in (96). This form of  $\operatorname{rcu}_s$  is more amenable to the saddlepoint approximation, since it is written as the tail probability of a sum of  $n+1$  independent variables,  $n$  of which are identically distributed.

Roughly speaking, we will approximate  $\operatorname{rcu}_s$  by expressing (130) in terms of the cumulant transform  $\kappa_n(\tau) = \log \mathbb{E}[e^{\tau \zeta_n}]$ , and then applying a suitable approximation to  $\kappa_n(\tau)$ . Since the cumulant transform is the logarithm of the Laplace transform of the probability density function  $f_{\zeta_n}(z)$  of  $\zeta_n$ , the density function itself is expressible as an inverse Laplace transform [32, Sec. 1.2], namely

$$f_{\zeta_n}(z) = \frac{1}{2\pi j} \int_{\hat{\tau}-j\infty}^{\hat{\tau}+j\infty} e^{\kappa_n(\tau) - \tau z} d\tau \quad (132)$$

for any  $\hat{\tau} > 0$  such that the integral converges. From (130),  $\operatorname{rcu}_s(n, M)$  is obtained by integrating (132) over  $z \in [0, \infty)$ . Interchanging the integration order, we obtain

$$\operatorname{rcu}_s(n, M) = \frac{1}{2\pi j} \int_{\hat{\tau}-j\infty}^{\hat{\tau}+j\infty} \int_0^\infty e^{\kappa_n(\tau) - \tau z} dz d\tau \quad (133)$$

$$= \frac{1}{2\pi j} \int_{\hat{\tau}-j\infty}^{\hat{\tau}+j\infty} e^{\kappa_n(\tau)} \frac{1}{\tau} d\tau. \quad (134)$$

Next, we compute the cumulant transform as

$$\kappa_n(\tau) = \log \mathbb{E}\left[e^{\tau \log \frac{M-1}{U} - \tau i_s^n(\mathbf{X}, \mathbf{Y})}\right] \quad (135)$$

$$= \tau \log(M-1) - \log(1-\tau) + \log \mathbb{E}\left[\left(\frac{\mathbb{E}[q^n(\overline{\mathbf{X}}, \mathbf{Y})^s | \mathbf{Y}]}{q^n(\mathbf{X}, \mathbf{Y})^s}\right)^\tau\right] \quad (136)$$

$$= \tau \log(M-1) - \log(1-\tau) - nE_0^{\text{iid}}(Q, \tau, s), \quad (137)$$



where (137) follows by expanding the expectation in (136) as a product from 1 to  $n$  and defining  $E_0^{\text{iid}}(Q, \tau, s)$  as in (54), with a fixed value of  $s > 0$  rather than a supremum. The second term in (137) is due to the expectation over  $U$ , and yields the condition  $\tau < 1$  for the convergence of the cumulant transform. Substituting (136) into (134) and replacing  $M$  by  $M + 1$ , we obtain

$$\text{rcu}_s(n, M + 1) = \frac{1}{2\pi j} \int_{\hat{\rho}-j\infty}^{\hat{\rho}+j\infty} e^{n(\rho R - E_0^{\text{iid}}(Q, \rho, s))} \frac{1}{\rho(1-\rho)} d\rho \quad (138)$$

where we have renamed  $\tau$  as  $\rho$ .

We choose  $\hat{\rho}$  to minimize the exponential term in the integrand in (138), namely

$$\hat{\rho} = \arg \max_{\rho \in [0, 1]} E_0^{\text{iid}}(Q, \rho, s) - \rho R, \quad (139)$$

where the restriction  $\rho \in [0, 1]$  arises from the above observation that  $\kappa_n(\tau)$  diverges when  $\tau$  is outside this range. While  $\kappa_n(\tau)$  also diverges when  $\tau = 0$  or  $\tau = 1$ , we can allow for these values by treating them as limiting cases. The dependence of  $\hat{\rho}$  on  $Q$ ,  $R$  and  $s$  is kept implicit. As discussed by Gallager, [1, Sec. 5.6],  $\hat{\rho}$  is a decreasing function of  $R$  which equals one for all rates below a critical rate, and approaches zero as  $R$  approaches the highest achievable rate, which is equal to  $I_s(Q)$  in (98) in the present setting.

The key step in the derivation is to approximate the exponential term in (138) by performing a Taylor expansion about  $\rho = \hat{\rho}$  in the exponent. Neglecting terms of order higher than two, we obtain

$$\rho R - E_0^{\text{iid}}(Q, \rho, s) \approx \hat{\rho} R - E_0^{\text{iid}}(Q, \hat{\rho}, s) + c_1(\rho - \hat{\rho}) + \frac{1}{2} c_2(\rho - \hat{\rho})^2, \quad (140)$$

where

$$c_1 \triangleq R - \left. \frac{\delta E_0^{\text{iid}}(Q, \rho, s)}{\delta \rho} \right|_{\rho=\hat{\rho}} \quad (141)$$

$$c_2 \triangleq - \left. \frac{\delta^2 E_0^{\text{iid}}(Q, \rho, s)}{\delta \rho^2} \right|_{\rho=\hat{\rho}}. \quad (142)$$

Applying the approximation (140) to (138), we obtain a quantity which can be written in terms of the error function and  $E_0^{\text{iid}}$ , and can thus be evaluated numerically with low complexity. Specifically, we use  $\frac{1}{\rho(1-\rho)} = \frac{1}{\rho} + \frac{1}{1-\rho}$ , apply the change of variables  $\rho' = \frac{\rho - \hat{\rho}}{j}$ , and evaluate the following integrals similarly to [32, Sec. 2.1]:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{jnc_1\rho' - \frac{1}{2}nc_2(\rho')^2} \frac{1}{\hat{\rho} + j\rho'} d\rho' = \frac{1}{2} \text{erfcx}_1 \left( \hat{\rho} \sqrt{\frac{nc_2}{2}}, c_1 \sqrt{\frac{n}{2c_2}} \right) \quad (143)$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{jnc_1\rho' - \frac{1}{2}nc_2(\rho')^2} \frac{1}{1 - \hat{\rho} - j\rho'} d\rho' = \frac{1}{2} \text{erfcx}_1 \left( (1 - \hat{\rho}) \sqrt{\frac{nc_2}{2}}, -c_1 \sqrt{\frac{n}{2c_2}} \right), \quad (144)$$

where  $\text{erfcx}_1(x, y) \triangleq \text{erfc}(x - y) \exp(x^2 - 2xy)$ . We thus obtain the saddlepoint approximation

$$\widehat{\text{rcu}}_s^{\text{iid}}(n, M) \triangleq \alpha^{\text{iid}}(\hat{\rho}, s) e^{-n(E_0^{\text{iid}}(Q, \hat{\rho}, s) - \hat{\rho} R)}, \quad (145)$$

where

$$\alpha^{\text{iid}}(\hat{\rho}, s) \triangleq \frac{1}{2} \text{erfcx}_1 \left( \hat{\rho} \sqrt{\frac{nc_2}{2}}, c_1 \sqrt{\frac{n}{2c_2}} \right) + \frac{1}{2} \text{erfcx}_1 \left( (1 - \hat{\rho}) \sqrt{\frac{nc_2}{2}}, -c_1 \sqrt{\frac{n}{2c_2}} \right). \quad (146)$$

The following theorem shows that  $\widehat{\text{rcu}}_s^{\text{iid}}$  is asymptotically equivalent to  $\text{rcu}_s^{\text{iid}}$  as  $n \rightarrow \infty$  for any rate  $R$  which yields a positive error exponent.

**Theorem 9.** Fix  $Q$ ,  $s > 0$  and  $R \in (0, I_s(Q))$ , where  $I_s(Q)$  is defined in (98). Let  $R_{\text{cr}}$  be the highest rate such that  $\hat{\rho} = 1$  for all  $R \leq R_{\text{cr}}$ . For the i.i.d. ensemble, we have the following: (i) If  $0 < R < R_{\text{cr}}$ , then

$$\lim_{n \rightarrow \infty} \frac{\widehat{\text{rcu}}_s^{\text{iid}}(n, e^{nR})}{\text{rcu}_s(n, e^{nR})} = 1 \quad (147)$$

and

$$\alpha^{\text{iid}}(\hat{\rho}, s) = 1 + o(1). \quad (148)$$

(ii) If  $R_{\text{cr}} < R < I_s(Q)$ , then (147) holds provided that  $i_s(X, Y)$  is a non-lattice variable. Furthermore, we have

$$\alpha^{\text{iid}}(\hat{\rho}, s) = \frac{1 + o(1)}{(1 - \hat{\rho})\hat{\rho}\sqrt{2\pi n c_2}} \quad (149)$$

regardless of whether  $i_s(X, Y)$  is a lattice variable.

*Proof:* See Appendix D. ■

The right hand sides of (148) and (149) yield precise estimates of  $\widehat{\text{rcu}}_s^{\text{iid}}$  as  $n \rightarrow \infty$ , but the coefficient in (146) is more accurate at finite values of  $n$ , particularly when  $\hat{\rho}$  is close to zero or one (i.e. values of  $R$  just below  $I_s(Q)$  or just above  $R_{\text{cr}}$ ). This observation is analogous to the comparison between (126) and (127) in Section VI-A.

The proof of the second part of Theorem 9 relies on  $i_s(X, Y)$  being a non-lattice variable. This is true for most decoding metrics and choices of  $Q$  and  $s$  in the non-binary and/or non-symmetric case. Having computed the approximation for several cases inducing a lattice variable, we have observed no noticeable difference in accuracy compared to the non-lattice case. In particular, the following example induces a lattice variable, and the approximations are still remarkably accurate.

We consider the example given in (40)–(41), using the parameters  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$ , and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . Under these parameters, we have  $I^{\text{GMI}}(Q) = 0.643$  and  $R_{\text{cr}} = 0.185$  bits/use. For the quantities for  $\text{rcu}_s$  and  $\widehat{\text{rcu}}_s$ , we choose the free parameter  $s$  to be the value which maximizes the error exponent at each rate. While this is not necessarily optimal at finite values of  $n$ , we will see that it yields good random-coding bounds. In Figure 4, we plot the error probability as a function of the rate with  $n = 45$ . We observe that  $\text{rcu}_s$  and  $\widehat{\text{rcu}}_s$  are nearly indistinguishable at all rates. Theorem 9 indicates that this should be the case for sufficiently large  $n$ , but it is interesting to observe that the same holds true even at small block lengths.

The curve  $e^{-nE_r^{\text{iid}}}$  approximates  $\text{rcu}_s$  accurately at low rates, but it is pessimistic at high rates. The normal approximation is obtained by rearranging (102) with the  $O(1)$  term omitted, using the value of  $s$  which achieves the GMI in (9). While its behavior is somewhat similar to that of  $\text{rcu}_s$ , but it is less precise than the saddlepoint approximation, and it is pessimistic at low rates. It is worth noting that the computational complexity of the saddlepoint approximation is similar to that of both the normal approximation and  $e^{-nE_r^{\text{iid}}}$ .

Our focus thus far has been on approximating  $\text{rcu}_s(n, M)$ . As shown in [33], one can apply similar techniques to approximate threshold-based random-coding bounds. The direct approximation of  $\text{rcu}(n, M)$  appears to be more

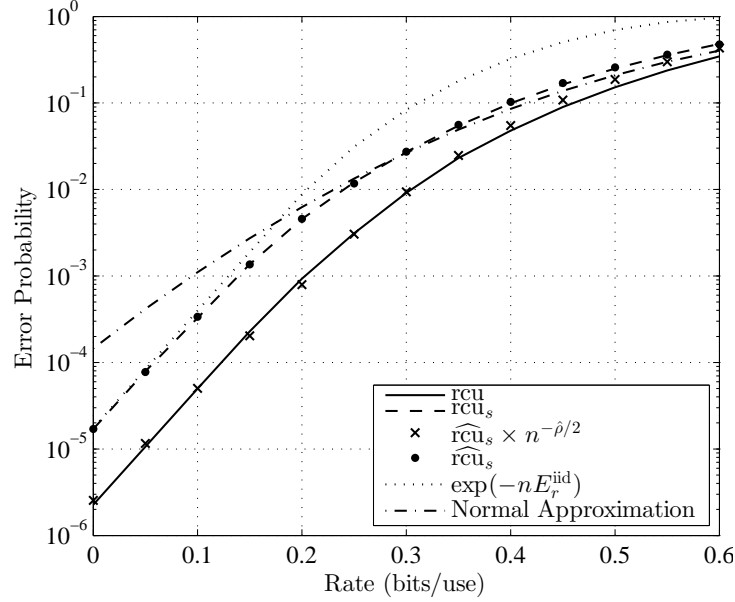


Figure 4. Random-coding bounds and saddlepoint approximation for the channel and metric defined in (40)–(41) using the i.i.d. ensemble. The parameters are  $n = 45$ ,  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$  and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ .

difficult, due to the conditional probability  $\mathbb{P}[q(\bar{\mathbf{X}}, \mathbf{y}) \geq q(\mathbf{x}, \mathbf{y})]$ . While  $\text{rcu}$  coincides with  $\text{rcu}_s$  for certain channels and metrics (e.g. binary-erasure channel (BEC) with ML decoding), the former is tighter in general. In particular, for sufficiently regular channels, the RCU bound with ML decoding achieves a pre-factor which is  $\Theta(n^{-\frac{1}{2}(1+\hat{\rho})})$  for rates above the critical rate [34, Thm. 1]. Theorem 9 gives a  $\Theta(n^{-\frac{1}{2}})$  pre-factor, which coincides with [34, Thm. 1] only for irregular channels resembling the BEC, and not for regular channels yielding the  $\Theta(n^{-\frac{1}{2}(1+\hat{\rho})})$  pre-factor. Based on this observation, we have plotted the quantity  $n^{-\frac{\hat{\rho}}{2}}\widehat{\text{rcu}}_s$  in Figure 4. Surprisingly, this quantity closely approximates  $\text{rcu}$ ; the deviation is most noticeable at high rates.

To facilitate the computation of  $\text{rcu}$  and  $\text{rcu}_s$  at larger block lengths, we consider the symmetric setup of  $\delta_0 = \delta_1 = \delta_2 = \delta = 0.1$ . Under these parameters, we have  $I(X; Y) = 0.633$  and  $R_{\text{cr}} = 0.192$  bits/use. In Figure 5, we plot the rate required for each random-coding bound and approximation to achieve a given error probability  $\epsilon = 10^{-8}$ , as a function of  $n$ . Once again the curves corresponding to  $\text{rcu}_s$  and  $\widehat{\text{rcu}}_s$  are nearly identical, and similarly for  $\text{rcu}$  and  $n^{-\frac{\hat{\rho}}{2}}\widehat{\text{rcu}}_s$ . The bound  $e^{-nE_r^{\text{iid}}}$  yields similar behavior to  $\text{rcu}_s$  at small block lengths, but the gap widens at larger block lengths.

In contrast to similar plots with larger choices of  $\epsilon$  (e.g. [18, Fig. 8]), the normal approximation is inaccurate, particularly at small block lengths. In the present example, the normal approximation is pessimistic, in the sense that it underestimates the achievable rate for a given  $(n, \epsilon)$  at all values of  $n$  plotted. However, in other cases the approximation can be optimistic. For example, changing the crossover probability to  $\delta = 0.01$ , the analogous plot to Figure 5 gives a normal approximation curve above that of  $\text{rcu}(n, M)$  at all block lengths shown. Further reducing the crossover probability to  $\delta = 0.001$ , the gap between  $\text{rcu}$  and the normal approximation becomes even

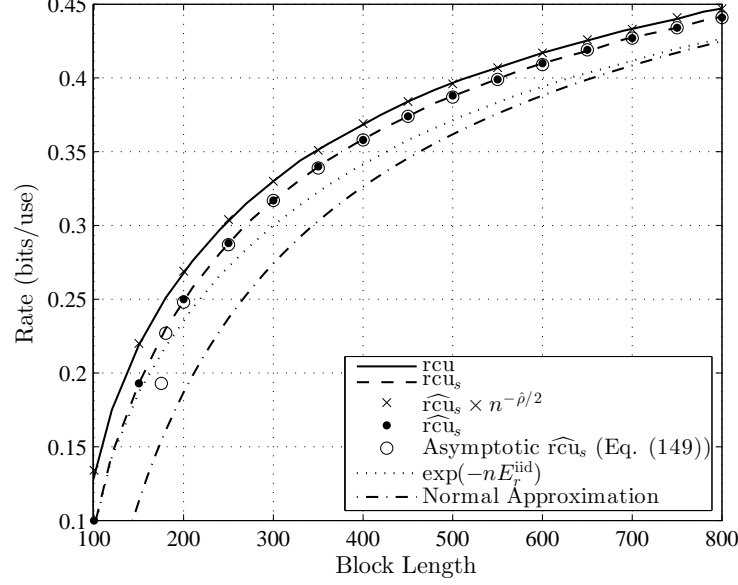


Figure 5. Rate required to achieve a target error probability  $\epsilon$  for the channel and metric defined in (40)–(41) using the i.i.d. ensemble. The parameters are  $\epsilon = 10^{-8}$ ,  $\delta_0 = \delta_1 = \delta_2 = \delta = 0.1$  and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ .

larger than that of Figure 5, but in the opposite direction. In contrast, the error exponent approximation for the i.i.d. ensemble is never optimistic in this sense, since it can be obtained by weakening  $\text{rcu}_s$ .

### C. Discussion

1) *Relation to the Normal Approximation:* In this subsection, we show that the normal approximation can be obtained by applying a limiting argument to the saddlepoint approximation as  $\hat{\rho} \rightarrow 0$ . Recall the definitions of  $I_s(Q)$  and  $U_s(Q)$  in (98)–(99). Consider a rate  $R$  which approaches  $I_s(Q)$ , so that  $\hat{\rho}$  in (139) tends to zero. We consider the Taylor expansion in (140) with zero in place of  $\hat{\rho}$ , yielding the approximation

$$\rho R - E_0^{\text{iid}}(Q, \rho, s) \approx (R - I_s(Q))\rho + \frac{1}{2}U_s(Q)\rho^2. \quad (150)$$

From (139), we expect  $\hat{\rho}$  to be approximately equal to the value obtained by letting the derivative of the right-hand side of (150) equal zero, i.e.  $\hat{\rho} \approx \frac{I_s(Q) - R}{U_s(Q)}$ . Within the same order of accuracy, we approximate  $c_2$  in (142) by the second derivative of (150), i.e.  $c_2 \approx U_s(Q)$ . The constant  $c_1$  does not need to be approximated, since it is precisely zero whenever  $\hat{\rho} \neq 1$ . Substituting these parameters into (145), we obtain

$$\widehat{\text{rcu}}_s^{\text{iid}}(n, M) \approx \frac{1}{2} \text{erfc} \left( (I_s(Q) - R) \sqrt{\frac{n}{2U_s(Q)}} \right). \quad (151)$$

Fixing  $\widehat{\text{rcu}}_s^{\text{iid}}(n, M)$  to a target value  $\epsilon$ , applying  $\text{erfc}(z) = 2Q(\sqrt{2}z)$ , and performing some simple algebra, we obtain an expression of the form (102), with the  $O(\log n)$  term omitted.

2) *Approximation for the Constant-Composition Ensemble:* Here we give a heuristic derivation of a saddlepoint approximation of  $\text{rcu}_s$  for the constant-composition ensemble. We do not prove a result analogous to Theorem 9, but we give a numerical example for which the approximation is accurate.

We introduce an arbitrary function  $a(x)$  and write  $a^n(\mathbf{x}) \triangleq \sum_{i=1}^n a(x_i)$ . We approximate  $\text{rcu}_s$  as follows:

$$\text{rcu}_s(n, M) = \mathbb{E} \left[ \min \left\{ 1, (M-1) \frac{\sum_{\bar{\mathbf{x}}} P_{\mathbf{X}}(\bar{\mathbf{x}}) q^n(\bar{\mathbf{x}}, \mathbf{Y})^s e^{a^n(\bar{\mathbf{x}})}}{q^n(\mathbf{X}, \mathbf{Y})^s e^{a^n(\mathbf{X})}} \right\} \right] \quad (152)$$

$$\approx \mathbb{E} \left[ \min \left\{ 1, M \frac{\sum_{\bar{\mathbf{x}}} Q^n(\bar{\mathbf{x}}) q^n(\bar{\mathbf{x}}, \mathbf{Y})^s e^{a^n(\bar{\mathbf{x}})}}{q^n(\mathbf{X}, \mathbf{Y})^s e^{a^n(\mathbf{X})}} \right\} \right] \quad (153)$$

$$= \mathbb{E} \left[ \min \left\{ 1, M e^{i_{s,a}^n(\mathbf{X}, \mathbf{Y})} \right\} \right] \quad (154)$$

$$= \mathbb{P} \left[ M e^{i_{s,a}^n(\mathbf{X}, \mathbf{Y})} \geq U \right] \quad (155)$$

$$= \mathbb{P} \left[ M e^{i_{s,a}^n(\mathbf{x}, \mathbf{Y})} \geq U \right]. \quad (156)$$

In (152) we have used the fact that  $a^n(\mathbf{x}) = a^n(\bar{\mathbf{x}})$  for any function  $a(\cdot)$  and codewords  $\mathbf{x}, \bar{\mathbf{x}}$  of the same type. In (153), we approximate the inner (but not outer) codeword distribution  $P_{\mathbf{X}}(\bar{\mathbf{x}})$  by the i.i.d. distribution  $Q^n(\bar{\mathbf{x}})$ , and we replace  $M-1$  by  $M$ . The third and fourth steps are identical to those for the i.i.d. ensemble in Section VI-B, and (156) holds for an arbitrary  $\mathbf{x} \in T^n(Q)$  by symmetry, where  $\mathbf{Y} \sim W^n(\cdot|\mathbf{x})$ .

The substitution of the i.i.d. ensemble in (153) is a heuristic step, and the approximation is pessimistic in the case that the cost function  $a(\cdot)$  is chosen suboptimally. However, based on numerical results, the approximation appears to be accurate when  $a(\cdot)$  is chosen to maximize  $E_0^{\text{cc}}(Q, \hat{\rho}, s, a)$  for the given value of  $s$ , where  $E_0^{\text{cc}}(Q, \rho, s, a)$  denotes the function  $E_0^{\text{cc}}$  in (55) with fixed values of  $s \geq 0$  and  $a(\cdot)$  rather than a supremum. By analyzing (156) in an identical fashion to the Section VI-B, we obtain precisely the approximation defined by (145)–(146), (139) and (141)–(142), with  $E_0^{\text{cc}}$  replacing  $E_0^{\text{iid}}$ . We denote this approximation by  $\widehat{\text{rcu}}_{s,a}^{\text{cc}}(n, M)$ .

Figure 6 plots the bounds and approximations for the constant-composition ensemble, again using the parameters  $n = 45$ ,  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$ , and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . For the quantities  $\text{rcu}_s$  and  $\widehat{\text{rcu}}_{s,a}$ , we choose  $s$  and  $a(\cdot)$  to maximize the error exponent at each rate. The normal approximation is obtained by rearranging (103) with the  $O(\log n)$  term omitted, using the values of  $s$  and  $a(\cdot)$  which achieve the LM rate in (4). Under these parameters, we have  $I^{\text{LM}}(Q) = 0.728$  and  $R_{\text{cr}} = 0.269$  bits/use. Despite the fact that its derivation was based on heuristic arguments, we see that  $\widehat{\text{rcu}}_{s,a}$  again approximates  $\text{rcu}_s$  remarkably well, and similarly for  $n^{-\frac{\hat{\rho}}{2}} \widehat{\text{rcu}}_{s,a}$  and  $\text{rcu}$ .

3) *Block Length as a Function of  $(R, \epsilon)$ :* The random-coding bounds in Section III are expressed as error probabilities in terms of the block length and number of messages. In order to find the required block length  $n$  for a given rate  $R$  and target error probability  $\epsilon$ , one can perform a bisection search [18]. An advantage of the error exponent approximation and normal approximation is that they can be inverted to find an explicit formula for  $n$  in terms of  $(R, \epsilon)$ . The error exponent approximation  $\bar{p}_e \approx e^{-n E_r^{\text{iid}}}$  yields the approximation

$$n \approx \frac{\log \epsilon}{E_r^{\text{iid}}(Q, R)}, \quad (157)$$

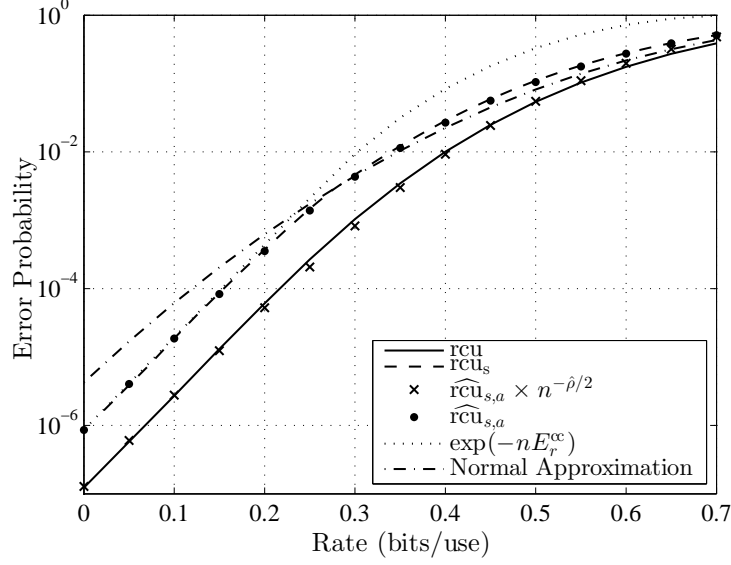


Figure 6. Random-coding bounds and saddlepoint approximation for the channel and metric defined in (40)–(41) using the constant-composition ensemble. The parameters are  $n = 45$ ,  $\delta_0 = 0.01$ ,  $\delta_1 = 0.05$ ,  $\delta_2 = 0.25$  and  $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ .

where  $R = \frac{1}{n} \log M$ . The normal approximation in (102) yields the approximation

$$n \approx \left( \frac{Q^{-1}(\epsilon) \sqrt{U_s(Q)}}{I_s(Q) - R} \right)^2. \quad (158)$$

While the approximation  $\widehat{r}_{cu_s}$  in (145) does not permit an explicit inversion, we can use the asymptotic expression in (149) to obtain an approximate expression for  $n$  at rates above the critical rate. Setting the quantity

$$\frac{1}{(1 - \hat{\rho}) \hat{\rho} \sqrt{2\pi n c_2}} e^{-n E_r^{\text{iid}}(Q, R)} \quad (159)$$

to a target value  $\epsilon$ , we obtain

$$n e^{2n E_r^{\text{iid}}(Q, R)} = \frac{1}{2\pi c_2 ((1 - \hat{\rho}) \hat{\rho})^2 \epsilon^2}. \quad (160)$$

Multiplying both sides by  $2E_r^{\text{iid}}$  and solving for  $n$ , we obtain the approximation

$$n \approx \frac{1}{2E_r^{\text{iid}}(Q, R)} \mathcal{W} \left( \frac{2E_r^{\text{iid}}(Q, R)}{2\pi c_2 ((1 - \hat{\rho}) \hat{\rho})^2 \epsilon^2} \right), \quad (161)$$

where  $\mathcal{W}(z)$  is the Lambert function, defined to be the value  $w$  such that  $we^w = z$ . As  $z \rightarrow \infty$ , we have  $\mathcal{W}(z) \approx \log z$ , and hence for small values of  $\epsilon$  we can further approximate (161) as

$$n \approx \frac{\log \epsilon}{E_r^{\text{iid}}(Q, R)} - \frac{1}{2E_r^{\text{iid}}(Q, R)} \log \frac{2E_r^{\text{iid}}(Q, R)}{2\pi c_2 ((1 - \hat{\rho}) \hat{\rho})^2}. \quad (162)$$

This can be seen as a refinement of (157).

Equations (161)–(162) should be used with care. From Theorem 9, we expect the resulting approximations to be accurate for any  $R \in (R_{\text{cr}}, I_s(Q))$  when the block length is large. However, (159) diverges as  $\hat{\rho} \rightarrow 0$  or  $\hat{\rho} \rightarrow 1$ , so for rates near  $R_{\text{cr}}$  or  $I_s(Q)$  one should instead use the approximation in (145), particularly when the block length

is small. The divergence as  $\hat{\rho} \rightarrow 1$  is seen in Figure 5, where the curve corresponding to (159) diverges from that of  $\widehat{rcu}_s$  when the corresponding rate is close to  $R_{cr} = 0.192$ . However, as the block length and rate increase, the two curves quickly become indistinguishable.

## VII. CONCLUSION AND DISCUSSION

Finite-length bounds on the random-coding error probability for mismatched decoders have been given. These bounds have been used to study the i.i.d. ensemble, constant-composition ensemble and cost-constrained ensembles. Error exponents and second-order coding rates have been given for each ensemble, and connections have been drawn between each. While the best performance in terms of rates and exponents is generally achieved by the constant-composition ensemble, we have shown that one can recover these gains using cost-constrained coding with at most two cost functions. Thus, cost-constrained coding is an attractive alternative to constant-composition coding; the two ensembles have similar performance, but only the latter can be directly applied in the case of infinite (possibly continuous) alphabets. Finally, saddlepoint approximations to the random-coding bounds have been presented. These have shown to provide accurate approximations to the random-coding bounds with low computational complexity.

Since the results in this paper are presented for a given input distribution  $Q$ , the extension to cost-constrained channels is straightforward. More concretely, suppose that each codeword  $\mathbf{x}$  is constrained to satisfy the constraint  $\frac{1}{n} \sum_{i=1}^n c(x_i) \leq \Gamma$  for some (system) cost function  $c(\cdot)$ . The i.i.d. ensemble is no longer suitable, since in all non-trivial cases it has a positive probability of producing codewords which violate the constraint. On the other hand, the results for the constant-composition ensemble remain unchanged provided that  $Q$  itself satisfies the cost constraint, i.e.  $\sum_x Q(x)c(x) \leq \Gamma$ .

For the cost-constrained ensemble, the extension is less trivial but still straightforward. The main change required is a modification of the definition of  $\mathcal{D}_n$  in (16) to include a constraint on the quantity  $\frac{1}{n} \sum_{i=1}^n c(x_i)$ . Unlike the pseudo-costs in (16), where the sample mean can be higher or lower than the true mean, the system cost of each codeword is constrained to be less than or equal to its mean. That is, the additional constraint is given by

$$\frac{1}{n} \sum_{i=1}^n c(x_i) \leq \phi_c \triangleq \sum_x Q(x)c(x), \quad (163)$$

or similarly with both upper and lower bounds (e.g.  $-\frac{\delta_c}{n} \leq \frac{1}{n} \sum_{i=1}^n c(x_i) - \phi_c \leq 0$  for some  $\delta_c$ ). Using this modified definition of  $\mathcal{D}_n$ , one can prove the subexponential behavior of  $\mu_n$  in Proposition 1, and the exponents and second-order rates for the cost-constrained ensemble remain valid under any  $Q$  such that  $\phi_c \leq \Gamma$ . In particular, one can recover the results of Theorems 7 and 8 using two pseudo-cost constraints and one system cost constraint in the definition of  $\mathcal{D}_n$ .

## APPENDIX

## A. Proof of Theorem 3

We begin by analyzing the cost-constrained ensemble. The codeword distribution in (15) can be written as

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\mu_n} \prod_{i=1}^n Q(x_i) \mathbf{1}\{\hat{P}_{\mathbf{x}} \in \mathcal{G}_n\}, \quad (164)$$

where  $\hat{P}_{\mathbf{x}}$  is the empirical distribution (type) of  $\mathbf{x}$ , and

$$\mathcal{G}_n \triangleq \left\{ P_X \in \mathcal{P}_n(\mathcal{X}) : |\mathbb{E}_P[a_l(X)] - \phi_l| \leq \frac{\delta}{n}, l = 1, \dots, L \right\}. \quad (165)$$

We define the sets

$$\mathcal{S}_n(\mathcal{G}_n) \triangleq \{P_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}) : P_X \in \mathcal{G}_n\} \quad (166)$$

$$\mathcal{T}_n(P_{XY}, \mathcal{G}_n) \triangleq \{\tilde{P}_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y}) : \tilde{P}_X \in \mathcal{G}_n, \tilde{P}_Y = P_Y, \mathbb{E}_{\tilde{P}}[\log q(X, Y)] \geq \mathbb{E}_P[\log q(X, Y)]\}. \quad (167)$$

Roughly speaking,  $\mathcal{S}_n$  is the set of possible types of  $(\mathbf{X}^{(m)}, \mathbf{Y})$  for the ensemble, and  $\mathcal{T}_n$  is the set of types of  $(\mathbf{X}^{(m')}, \mathbf{Y})$  which lead to errors given that  $(\mathbf{X}^{(m)}, \mathbf{Y}) \in T^n(P_{XY})$ , where  $m$  is the transmitted message and  $m'$  is a different message.

We have from (25) and (28) that  $\bar{p}_e \doteq \text{rcu}(n, M)$ . Expanding  $\text{rcu}(n, M)$  in terms of types, we obtain

$$\bar{p}_e(n, M) \doteq \sum_{P_{XY} \in \mathcal{S}_n(\mathcal{G}_n)} \mathbb{P}[(\mathbf{X}, \mathbf{Y}) \in T^n(P_{XY})] \psi(P_{XY}), \quad (168)$$

where

$$\psi(P_{XY}) \triangleq \min \left\{ 1, (M-1) \sum_{\tilde{P}_{XY} \in \mathcal{T}_n(P_{XY}, \mathcal{G}_n)} \mathbb{P}[(\bar{\mathbf{X}}, \mathbf{y}) \in T^n(\tilde{P}_{XY})] \right\} \quad (169)$$

and  $\mathbf{y}$  is an arbitrary sequence with type  $P_Y$ . From (164), the distribution of  $\bar{\mathbf{X}}$  is the same as that of  $\mathbf{X}' \sim \prod_{i=1}^n Q(x_i)$  conditioned on the event that  $\mathbf{X}' \in \mathcal{G}_n$ . From Proposition 1, the normalizing constant in (164) satisfies  $\mu_n \doteq 1$ , and thus

$$\psi(P_{XY}) \doteq \min \left\{ 1, (M-1) \sum_{\tilde{P}_{XY} \in \mathcal{T}_n(P_{XY}, \mathcal{G}_n)} \mathbb{P}[(\mathbf{X}', \mathbf{y}) \in T^n(\tilde{P}_{XY})] \right\} \quad (170)$$

$$\doteq \max_{\tilde{P}_{XY} \in \mathcal{T}_n(P_{XY}, \mathcal{G}_n)} \min \left\{ 1, (M-1) \exp(-nD(\tilde{P}_{XY} \| Q \times \tilde{P}_Y)) \right\}, \quad (171)$$

where (171) follows from the property of types in [15, Eq. (18)], and the fact that the number of joint types is polynomial in  $n$ . Applying a nearly identical argument to the probability in (168), we obtain

$$\bar{p}_e(n, M) \doteq \max_{P_{XY} \in \mathcal{S}_n(\mathcal{G}_n)} \exp(-nD(P_{XY} \| Q \times W)) \psi(P_{XY}). \quad (172)$$

From the above-mentioned properties of types, the implied subexponential factor in (171) can be bounded in both directions by polynomials which do not depend on  $P_{XY}$ . We can therefore substitute (171) into (172), yielding

$$\bar{p}_e(n, e^{nR}) \doteq \exp(-nE_{r,n}(Q, R, \mathcal{G}_n)), \quad (173)$$



where

$$E_{r,n}(Q, R, \mathcal{G}_n) \triangleq \min_{P_{XY} \in \mathcal{S}_n(\mathcal{G}_n)} \min_{\tilde{P}_{XY} \in \mathcal{T}_n(P_{XY}, \mathcal{G}_n)} D(P_{XY} \| Q \times W) + \left[ D(\tilde{P}_{XY} \| Q \times \tilde{P}_Y) - R \right]^+. \quad (174)$$

To conclude the proof of (50), we need to show that the constraints  $|\mathbb{E}_P[a_l(X)] - \phi_l| \leq \frac{\delta}{n}$  in (165) can be replaced by  $\mathbb{E}_P[a_l(X)] = \phi_l$ , regardless of the value of  $\delta$ . This follows using the same argument as that used to replace minimizations over types by minimizations over all distributions [15]; a simple continuity argument shows that the exponential behavior is unaffected.

The i.i.d. exponent in (48) follows from (50) by setting  $L = 0$ . The constant-composition exponent in (49) follows by treating the constant-composition ensemble as a special case of the cost-constrained ensemble, using the parameters given in Section II. Due to the constraints  $P_X = Q$  and  $\tilde{P}_X = P_X$  in (43) and (46), the quantity  $D(\tilde{P}_{XY} \| Q \times \tilde{P}_Y)$  in (50) can be replaced by  $I_{\tilde{P}}(X; Y)$ .

#### B. Proof of Theorem 4

Using  $[\alpha]^+ = \max_{\rho \in [0,1]} \rho\alpha$ , the convexity properties of the optimization problem, and Fan's minimax theorem [35], the expression in (49) can be written as

$$E_r^{\text{cc}}(Q, R) = \max_{\rho \in [0,1]} \hat{E}_0^{\text{cc}}(Q, \rho) - \rho R \quad (175)$$

where

$$\hat{E}_0^{\text{cc}}(Q, \rho) \triangleq \min_{P_{XY} \in \mathcal{S}^{\text{cc}}(Q)} \min_{\tilde{P}_{XY} \in \mathcal{T}^{\text{cc}}(P_{XY})} D(P_{XY} \| Q \times W) + \rho I_{\tilde{P}}(X; Y). \quad (176)$$

It remains to show that  $\hat{E}_0^{\text{cc}}(Q, \rho) = E_0^{\text{cc}}(Q, \rho)$ . We will show this by considering the minimizations in (176) one at a time, and using Lagrange duality. In fact, we can immediately recognize

$$\min_{\tilde{P}_{XY} \in \mathcal{T}(P_{XY})} I_{\tilde{P}}(X; Y). \quad (177)$$

as the LM rate for the channel  $P_{Y|X}$  with input distribution  $P_X$ ; see (5). Thus, replacing this minimization with the equivalent dual expression in (4), we conclude that the minimization in (176) is equivalent to

$$\min_{P_{XY} \in \mathcal{S}^{\text{cc}}(Q)} \sup_{s \geq 0, a(\cdot)} \sum_{x,y} P_{XY}(x, y) \left( \log \frac{P_{XY}(x, y)}{Q(x)W(y|x)} - \rho \log \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right). \quad (178)$$

Using Fan's minimax theorem [35], we can swap the order of the minimum and the supremum. Hence, we first minimize the objective in (178) over  $P_{XY} \in \mathcal{S}(Q)$  with  $s$  and  $a(x)$  fixed. Introducing a Lagrange multiplier  $\eta(x)$  corresponding to the constraint  $P_X = Q$ , the Lagrangian is given by

$$L_0 = \sum_{x,y} P_{XY}(x, y) \left( \log \frac{P_{XY}(x, y)}{Q(x)W(y|x)} - \rho \log \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} - \eta(x) \right) + \sum_x \eta(x) Q(x). \quad (179)$$

Setting  $\frac{\delta L_0}{\delta P_{XY}(x, y)} = 0$  gives

$$1 + \log \frac{P_{XY}(x, y)}{Q(x)W(y|x)} - \rho \log \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} - \eta(x) = 0. \quad (180)$$

Solving (180) for  $P_{XY}(x, y)$  gives

$$P_{XY}(x, y) = Q(x)W(y|x)e^{-1+\eta(x)} \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right)^\rho. \quad (181)$$

Applying the constraint  $P_X = Q$  to (181), we obtain

$$\sum_y W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right)^\rho = e^{1-\eta(x)}, \quad (182)$$

which implies

$$\eta(x) = 1 - \log \sum_y W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right)^\rho. \quad (183)$$

Substituting (180) into (179) gives the simplified Lagrangian expression  $L_0 = -1 + \sum_x Q(x)\eta(x)$ , and applying (183), we obtain

$$L_0 = - \sum_x Q(x) \log \sum_y W(y|x) \left( \frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x}, y)^s e^{a(\bar{x})}}{q(x, y)^s e^{a(x)}} \right)^\rho. \quad (184)$$

Since the optimization under consideration is a convex optimization problem, the duality gap is zero [11]. The proof is concluded by taking the supremum over  $(s, a)$ .

### C. Proof of Theorem 5

It suffices to prove the theorem for  $\tilde{Q}$  with the same support as  $Q$ , since otherwise the divergence in (82) is infinite and  $\tilde{Q}$  cannot achieve the maximum. We fix  $\tilde{Q}(x)$  such that  $\text{supp}(\tilde{Q}) = \text{supp}(Q)$ , and define  $\tilde{a}(x)$  such that

$$e^{a(x)} \frac{Q(x)}{\tilde{Q}(x)} = e^{\tilde{a}(x)}. \quad (185)$$

Substituting (185) into (55) yields

$$E_0^{\text{cc}}(Q, \rho) = \sup_{s \geq 0, \tilde{a}(\cdot)} - \sum_x Q(x) \log \sum_y \left( \frac{Q(x)}{\tilde{Q}(x)} \right)^\rho W(y|x) \left( \frac{\sum_{\bar{x}} \tilde{Q}(\bar{x})q(\bar{x}, y)^s e^{\tilde{a}(\bar{x})}}{q(x, y)^s e^{\tilde{a}(x)}} \right)^\rho \quad (186)$$

$$= \sup_{s \geq 0, \tilde{a}(\cdot)} - \sum_x Q(x) \log \sum_y \frac{\tilde{Q}(x)}{Q(x)} W(y|x) \left( \frac{\sum_{\bar{x}} \tilde{Q}(\bar{x})q(\bar{x}, y)^s e^{\tilde{a}(\bar{x})}}{q(x, y)^s e^{\tilde{a}(x)}} \right)^\rho - (1 + \rho)D(Q\|\tilde{Q}) \quad (187)$$

$$\geq \sup_{s \geq 0, \tilde{a}(\cdot)} - \log \sum_{x, y} \tilde{Q}(x) W(y|x) \left( \frac{\sum_{\bar{x}} \tilde{Q}(\bar{x})q(\bar{x}, y)^s e^{\tilde{a}(\bar{x})}}{q(x, y)^s e^{\tilde{a}(x)}} \right)^\rho - (1 + \rho)D(Q\|\tilde{Q}), \quad (188)$$

where (186) follows since, for any given  $(Q, \tilde{Q})$  pair, maximizing over  $a(\cdot)$  is equivalent to maximizing over  $\tilde{a}(\cdot)$ , and (188) follows from Jensen's inequality. It remains to show that  $\tilde{Q}(x)$  (188) holds with equality when  $Q$  is chosen to maximize (187). It suffices to show that the same holds true when the supremum over  $s$  on each side is replaced by any given value of  $s$ . We will show that equality holds in (188) when  $\tilde{Q}$  and  $\tilde{a}(\cdot)$  are chosen to maximize the objective in (188), and then show that the resulting  $a(\cdot)$  (which is computed from (185)) maximizes the objective in (55) (with respect to  $a_1(\cdot)$ ) for the given value of  $s$ . We define

$$f(x, \tilde{Q}, \tilde{a}) \triangleq \sum_y W(y|x) q(x, y)^{-\rho s} \left( \sum_{\bar{x}} \tilde{Q}(\bar{x}) q(\bar{x}, y)^s e^{\tilde{a}(\bar{x})} \right)^\rho. \quad (189)$$

The optimization problems corresponding to (55) and (188) can respectively be written as

$$\underset{a(\cdot)}{\text{maximize}} - \sum_x Q(x) \left( -\rho a(x) + \log f(x, Q, a) \right) \quad (190)$$

$$\underset{\tilde{Q}, \tilde{a}(\cdot)}{\text{maximize}} - \log \sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a}) - (1 + \rho) D(Q \| \tilde{Q}). \quad (191)$$

To obtain the KKT conditions, the following derivatives will be useful.

$$\frac{\delta f(x, \tilde{Q}, \tilde{a})}{\delta \tilde{Q}(x')} = \sum_y W(y|x) q(x, y)^{-\rho s} \rho \left( \sum_{\bar{x}} \tilde{Q}(\bar{x}) q(\bar{x}, y)^s e^{\tilde{a}(\bar{x})} \right)^{\rho-1} q(x', y)^s e^{\tilde{a}(x')} \triangleq g(x, x', \tilde{Q}, \tilde{a}) \quad (192)$$

$$\frac{\delta f(x, \tilde{Q}, \tilde{a})}{\delta a(x')} = \tilde{Q}(x') g(x, x', \tilde{Q}, \tilde{a}) \quad (193)$$

The Lagrangians corresponding to (190) and (191) are respectively given by

$$L_1 = - \sum_x Q(x) \left( -\rho a(x) + \log f(x, Q, a) \right) \quad (194)$$

$$L_2 = - \log \sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a}) - (1 + \rho) \sum_x Q(x) \log \frac{Q(x)}{\tilde{Q}(x)} + \mu \left( \sum_x \tilde{Q}(x) - 1 \right), \quad (195)$$

where  $\mu$  is a Lagrange multiplier.

We begin by analyzing the KKT conditions of (191). While the objective is non-concave in  $(\tilde{Q}, \tilde{a})$ , the KKT conditions are still necessary for optimality. However, we must rule out the possibility that the maximization is obtained by a limiting value, e.g. as  $|\tilde{a}(x)|$  grows unbounded for some  $x$ . We first observe that each  $\tilde{Q}(x)$  must remain bounded away from zero for  $x \in \text{supp}(Q)$ . To see this, we note that (i) the first term in (191) can easily be upper bounded by a finite constant, and (ii) as  $\tilde{Q}(x) \rightarrow 0$  for  $x \in \text{supp}(Q)$ , the divergence term in (191) tends toward  $-\infty$ . Secondly, since the objective is unchanged if the same constant value is added to each  $a(x)$  (cf. (188)), it suffices to fix the highest value of  $\tilde{a}(x)$  to zero and show that the maximum cannot be achieved in the limit as the remaining values approach  $-\infty$ . Indeed, since our setup assumes  $q(x, y) > 0$  wherever  $W(y|x) > 0$  (cf. Section I-A),  $\tilde{a}(x) \rightarrow -\infty$  would imply a division by a vanishing quantity in (188), yielding an objective of  $-\infty$ .

Setting  $\frac{\delta L_2}{\delta \tilde{Q}(x')} = 0$  gives

$$- \frac{e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a}) + \sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} g(x, x', \tilde{Q}, \tilde{a})}{\sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a})} + (1 + \rho) \frac{Q(x')}{\tilde{Q}(x')} + \mu = 0. \quad (196)$$

Setting  $\frac{\delta L_2}{\delta \tilde{a}(x')} = 0$  gives

$$\frac{-1}{\sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a})} \left( -\tilde{Q}(x') \rho e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a}) + \sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} \tilde{Q}(x') g(x, x', \tilde{Q}, \tilde{a}) \right) = 0, \quad (197)$$

and hence

$$\frac{\sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} g(x, x', \tilde{Q}, \tilde{a})}{e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a})} = \rho. \quad (198)$$

Substituting (198) into (196) gives

$$- (1 + \rho) \frac{e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a})}{\sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a})} + (1 + \rho) \frac{Q(x')}{\tilde{Q}(x')} + \mu = 0. \quad (199)$$

Multiplying both sides by  $\tilde{Q}(x')$  and summing over  $x'$  gives  $\mu = 0$ , and hence

$$\frac{\frac{\tilde{Q}(x')}{Q(x')} e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a})}{\sum_x \tilde{Q}(x) e^{-\rho \tilde{a}(x)} f(x, \tilde{Q}, \tilde{a})} = 1 \quad (200)$$

for all  $x'$ . It follows from (200) that the quantity

$$\frac{\tilde{Q}(x')}{Q(x')} e^{-\rho \tilde{a}(x')} f(x', \tilde{Q}, \tilde{a}) \quad (201)$$

is independent of  $x'$ , which is a sufficient condition for Jensen's inequality to hold with equality in (188).

We now turn to the KKT conditions for (55). We will show that if  $(\tilde{Q}, \tilde{a})$  satisfies the KKT conditions for (191), then the cost function  $a(\cdot)$  which is computed using (185) satisfies the KKT conditions for (190). We denote such a  $(\tilde{Q}, \tilde{a})$  pair by  $(\tilde{Q}^*, \tilde{a}^*)$ , and we denote the corresponding  $a(\cdot)$  by  $a^*(\cdot)$ .

Setting  $\frac{\delta L_1}{\delta a(x')} = 0$  gives

$$\rho Q(x') - \sum_x \frac{Q(x) Q(x') g(x, x', Q, a)}{f(x, Q, a)} = 0, \quad (202)$$

or equivalently

$$\sum_x \frac{Q(x) g(x, x', Q, a)}{f(x, Q, a)} = \rho. \quad (203)$$

We must show that  $a^*(\cdot)$  satisfies (203) for all  $x'$ ; this will complete the proof since (190) is a convex optimization problem with affine constraints, and hence the KKT conditions are both necessary and sufficient [11]. Using (185) and the definitions of  $f$  and  $g$ , we have

$$f(x, Q, a^*) = f(x, \tilde{Q}^*, \tilde{a}^*) \quad (204)$$

$$g(x, x', Q, a^*) = \frac{\tilde{Q}^*(x')}{Q(x')} g(x, x', \tilde{Q}^*, \tilde{a}^*). \quad (205)$$

Thus,  $a^*(\cdot)$  satisfies (203) if and only if  $(\tilde{Q}^*, \tilde{a}^*)$  satisfies

$$\frac{\tilde{Q}^*(x')}{Q(x')} \sum_x \frac{Q(x) g(x, x', \tilde{Q}^*, \tilde{a}^*)}{f(x, \tilde{Q}^*, \tilde{a}^*)} = \rho, \quad (206)$$

or equivalently

$$\frac{\tilde{Q}^*(x')}{Q(x')} \sum_x \frac{\tilde{Q}^*(x) e^{-\rho \tilde{a}^*(x)} g(x, x', \tilde{Q}^*, \tilde{a}^*)}{\frac{\tilde{Q}^*(x)}{Q(x)} e^{-\rho \tilde{a}^*(x)} f(x, \tilde{Q}^*, \tilde{a}^*)} = \rho. \quad (207)$$

Since the quantity in (201) is independent of  $x'$  under  $(\tilde{Q}^*, \tilde{a}^*)$ , it follows that (207) holds if and only if

$$\frac{\tilde{Q}^*(x')}{Q(x')} \sum_x \frac{\tilde{Q}^*(x) e^{-\rho \tilde{a}^*(x)} g(x, x', \tilde{Q}^*, \tilde{a}^*)}{\frac{\tilde{Q}^*(x')}{Q(x')} e^{-\rho \tilde{a}^*(x')} f(x', \tilde{Q}^*, \tilde{a}^*)} = \rho \quad (208)$$

or equivalently

$$\frac{\sum_x \tilde{Q}^*(x) e^{-\rho \tilde{a}^*(x)} g(x, x', \tilde{Q}^*, \tilde{a}^*)}{e^{-\rho \tilde{a}^*(x')} f(x', \tilde{Q}^*, \tilde{a}^*)} = \rho. \quad (209)$$

Finally, since (209) coincides with (198), which in turn was a result of the KKT conditions for (191), it follows that (209) holds, and hence  $a^*(\cdot)$  satisfies (203) for all  $x'$ . Thus,  $a^*(\cdot)$  maximizes the objective in (190).

#### D. Proof of Theorem 9

The proof of the second part of the theorem is based on the analysis of Bahadur and Rao [36], while the proof of the first part is more straightforward. In both cases, it is easy to show that the use of  $M$  in place of  $M - 1$  in the definition of  $\zeta_n$  does not affect the statement of the theorem. Thus, throughout the proof we focus on the bound  $\text{rcu}_s(n, M + 1)$ .

1) *Below the Critical Rate:* Using Gallager's properties of error exponents [1, Sec. 5.6], we have  $\hat{\rho} = 1$ ,  $c_1 < 0$  and  $c_2 > 0$  for all  $R < R_{\text{cr}}$ , and an asymptotic expansion of  $\alpha^{\text{iid}}(\hat{\rho}, s)$  as  $n \rightarrow \infty$  yields (148). To prove (147), we need to show that

$$\text{rcu}_s(n, M + 1) = (1 + o(1)) \exp \left( n(E_0^{\text{iid}}(Q, 1, s) - R) \right). \quad (210)$$

The upper bound follows immediately by applying  $\min\{1, \alpha\} \leq \alpha$  to (26). To obtain a matching lower bound, we write

$$\text{rcu}_s(n, M + 1) = \mathbb{P} \left[ i_s^n(\mathbf{X}, \mathbf{Y}) \leq \log M \right] + M \mathbb{E} \left[ e^{-i_s^n(\mathbf{X}, \mathbf{Y})} \mathbb{1} \{ i_s^n(\mathbf{X}, \mathbf{Y}) > \log M \} \right] \quad (211)$$

$$= \mathbb{P} \left[ i_s^n(\mathbf{X}, \mathbf{Y}) \leq \log M \right] + M \mathbb{E} \left[ e^{-i_s^n(\mathbf{X}, \mathbf{Y})} \right] - M \mathbb{E} \left[ e^{-i_s^n(\mathbf{X}, \mathbf{Y})} \mathbb{1} \{ i_s^n(\mathbf{X}, \mathbf{Y}) \leq \log M \} \right] \quad (212)$$

$$\geq \mathbb{P} \left[ i_s^n(\mathbf{X}, \mathbf{Y}) \leq \log M \right] + M \mathbb{E} \left[ e^{-i_s^n(\mathbf{X}, \mathbf{Y})} \right] - M^{1+\rho'} \mathbb{E} \left[ e^{-(1+\rho')i_s^n(\mathbf{X}, \mathbf{Y})} \right] \quad (213)$$

where (211) follows by splitting the  $\min\{1, \cdot\}$  in (26), (212) follows since  $\mathbb{1}\{A\} + \mathbb{1}\{A^c\} = 1$ , and (213) follows for any  $\rho' \geq 0$  by upper bounding the indicator function.

The second term in the summation in (213) is simply the right-hand side of (210) with the  $o(1)$  term omitted, so it only remains to show that the other two terms decay at a faster rate. Using the strong large deviations result of [36], the first term decays with an exponent at least as high as  $E_0^{\text{iid}}(Q, 1, s) - R$ , with an additional  $\frac{1}{\sqrt{n}}$  pre-factor. Furthermore, since the sphere-packing bound exceeds the random-coding bound below the critical rate [1, Sec. 5], there exists a  $\rho'$  which makes the third term in (213) decay with an exponent exceeding  $E_0^{\text{iid}}(Q, 1, s) - R$ .

2) *Above the Critical Rate:* Using Gallager's properties of error exponents [1, Sec. 5.6], we have  $\hat{\rho} = 1$ ,  $c_1 = 0$  and  $c_2 > 0$  for all  $R > R_{\text{cr}}$ , and an asymptotic expansion of  $\alpha^{\text{iid}}(\hat{\rho}, s)$  as  $n \rightarrow \infty$  yields (149). To prove (147), we combine the techniques of [36] with the refined version of the central limit theorem given in [16, Sec. XVI.4, Thm. 1].

Using the expression for  $\text{rcu}_s$  in (130), we have

$$\text{rcu}_s(n, M + 1) = \mathbb{P} \left[ nR - \sum_{i=1}^n i_s(X_i, Y_i) \geq \log U \right]. \quad (214)$$

This expression resembles the upper tail probability of an i.i.d. sum of random variables, for which asymptotic estimates were given by Bahadur and Rao [36]. Due to the presence of the  $\log U$  term in (214), the results of [36] are not directly applicable. However, by performing a similar analysis, we will obtain the desired result.

Let  $F(t)$  denote the cumulative distribution function (CDF) of  $R - i_s(X, Y)$ , and let  $Z_1, \dots, Z_n$  be distributed according to the tilted CDF

$$F_Z(z) = e^{-E_r^{\text{iid}}} \int_{-\infty}^z e^{\hat{\rho}t} dF(t) \quad (215)$$

where we write  $E_r^{\text{iid}}$  as a shorthand for  $\exp(E_0^{\text{iid}}(Q, \hat{\rho}, s) - \hat{\rho}R)$ . Using an identical argument to [36, Lemma 2], we can write

$$\text{rcu}_s(n, M+1) = I_n e^{-nE_r^{\text{iid}}}, \quad (216)$$

where

$$I_n \triangleq \mathbb{E} \left[ e^{-\hat{\rho} \sum_i Z_i} \mathbb{1} \left\{ \sum_i Z_i \geq \log U \right\} \right]. \quad (217)$$

Using an integration by parts, we obtain [36]

$$I_n = \int_0^1 \hat{\rho} \sigma \sqrt{n} \int_{\frac{\log u}{\sigma \sqrt{n}}}^{\infty} e^{-\hat{\rho} \sigma \sqrt{n} z} \left( F_n(z) - F_n \left( \frac{\log u}{\sigma \sqrt{n}} \right) \right) dz du, \quad (218)$$

where  $\sigma^2 \triangleq \text{Var}[Z]$ , and  $F_n(z)$  is the CDF of  $\frac{\sum_i Z_i}{\sigma \sqrt{n}}$ .

Let  $\Phi(z)$  denote the CDF of a zero-mean unit-variance Gaussian random variable. From [16, Sec. XVI.4, Thm. 1], we have

$$F_n(z) = \Phi(z) + G_n(z) + \tilde{F}_n(z), \quad (219)$$

where  $\tilde{F}_n(z) = o(n^{-\frac{1}{2}})$  uniformly in  $z$ , and

$$G_n(z) \triangleq \frac{K}{\sqrt{n}} (1 - z^2) e^{-\frac{1}{2} z^2} \quad (220)$$

for some constant  $K$  depending only on the second and third moments of  $Z$ , both of which are finite for any DMC. Substituting (219) into (218), we obtain

$$I_n = I_{1,n} + I_{2,n} + o\left(\frac{1}{\sqrt{n}}\right), \quad (221)$$

where

$$I_{1,n} \triangleq \int_0^1 \hat{\rho} \sigma \sqrt{n} \int_{\frac{\log u}{\sigma \sqrt{n}}}^{\infty} e^{-\hat{\rho} \sigma \sqrt{n} z} \left( \Phi_n(z) - \Phi_n \left( \frac{\log u}{\sigma \sqrt{n}} \right) \right) dz du, \quad (222)$$

$$I_{2,n} \triangleq \int_0^1 \hat{\rho} \sigma \sqrt{n} \int_{\frac{\log u}{\sigma \sqrt{n}}}^{\infty} e^{-\hat{\rho} \sigma \sqrt{n} z} \left( G_n(z) - G_n \left( \frac{\log u}{\sigma \sqrt{n}} \right) \right) dz du. \quad (223)$$

To evaluate (222) we use integration by parts to obtain an expression of the same form as (217), with each  $Z_i$  replaced by a zero-mean random variable with variance  $\sigma^2$ . A direct evaluation of the resulting expectation yields precisely the pre-factor in (146), with 0 and  $\sigma^2$  in place of  $c_1$  and  $c_2$  respectively. Evaluating (141) and (142) directly, and using the expression for  $\sigma^2$  described by [36, Eqs. (4),(8)], it is easily verified that these coefficients coincide. From (149) and (221), it only remains to show that  $I_{2,n} = o\left(\frac{1}{\sqrt{n}}\right)$ .

Following [36, Eq. (28)], we can use integration by parts to write

$$I_{2,n} = \frac{1}{2\pi} \int_0^1 \int_{\frac{\log u}{\sigma \sqrt{n}}}^{\infty} e^{-\hat{\rho} \sigma \sqrt{n} z} G'_n(z) dz du, \quad (224)$$

where  $G'_n(z)$  is the derivative of  $G_n(z)$ , namely

$$G'_n(z) = \frac{K}{\sqrt{n}} (z^3 - 3z) e^{-\frac{1}{2} z^2}. \quad (225)$$

Changing the order of integration in (224), we obtain

$$I_{2,n} = \frac{K}{2\pi\sqrt{n}} \int_{-\infty}^{\infty} \int_0^{\min\{1, e^{\sigma\sqrt{n}z}\}} e^{-\hat{\rho}\sigma\sqrt{n}z} (z^3 - 3z) e^{-\frac{1}{2}z^2} du dz \quad (226)$$

$$= \frac{K}{2\pi\sqrt{n}} \int_{-\infty}^{\infty} e^{-\hat{\rho}\sigma\sqrt{n}z} (z^3 - 3z) \min\{1, e^{\sigma\sqrt{n}z}\} e^{-\frac{1}{2}z^2} dz \quad (227)$$

$$= \frac{K}{2\pi\sqrt{n}} \left( \int_0^{\infty} e^{-\hat{\rho}\sigma\sqrt{n}z} (z^3 - 3z) e^{-\frac{1}{2}z^2} dz + \int_{-\infty}^0 e^{(1-\hat{\rho})\sigma\sqrt{n}z} (z^3 - 3z) e^{-\frac{1}{2}z^2} dz \right) \quad (228)$$

Evaluating the integrals in (228), we conclude that both are  $O(n^{-1})$ , and hence  $I_{2,n} = O(n^{-\frac{3}{2}})$ . Thus, using (221), we conclude that  $I_n = I_{1,n} + o(n^{-\frac{1}{2}})$ . Since  $I_{1,n}$  is the desired pre-factor, and decays as  $\frac{1}{\sqrt{n}}$  (cf. (149)), we conclude that  $\lim_{n \rightarrow \infty} \frac{I_n}{I_{1,n}} = 1$ , thus concluding the proof.

## REFERENCES

- [1] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.
- [2] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *Arch. Elek. Über.*, vol. 47, no. 4, pp. 228–239, 1993.
- [3] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 35–43, Jan. 1995.
- [4] J. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, MIT, 1983.
- [5] V. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1889–1902, Nov. 1995.
- [6] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.
- [7] A. Ganti, A. Lapidoth, and E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.
- [8] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep. 1996.
- [9] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.
- [10] T. Fischer, "Some remarks on the role of inaccuracy in Shannon's theory of information transmission," in *Trans. 8th Prague Conf. on Inf. Theory*, 1971, pp. 211–226.
- [11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [12] S. Shamai and I. Sason, "Variations on the Gallager bounds, connections, and applications," *IEEE Trans. Inf. Theory*, vol. 48, no. 12, pp. 3029–3051, Dec. 2002.
- [13] A. Nazari, A. Anastasopoulos, and S. S. Pradhan, "Error exponent for multiple-access channels: Lower bounds," arXiv:1010.1303v1 [cs.IT].
- [14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.
- [15] R. Gallager, "Fixed composition arguments and lower bounds to error probability," <http://web.mit.edu/gallager/www/notes/notes5.pdf>.
- [16] W. Feller, *An introduction to probability theory and its applications*, 2nd ed. John Wiley & Sons, 1971, vol. 2.
- [17] P. S. Griffin, "Matrix normalized sums of independent identically distributed random vectors," *Annals of Probability*, vol. 14, no. 1, pp. 224–246, 1986.
- [18] Y. Polyanskiy, V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [19] D. de Caen, "A lower bound on the probability of a union," *Discrete Mathematics*, vol. 169, pp. 217–220, 1997.
- [20] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Ensemble-tight error exponents for mismatched decoders," in *Allerton Conf. on Comm., Control and Comp.*, Monticello, IL, Oct. 2012.
- [21] R. Fano, *Transmission of information: A statistical theory of communications*. MIT Press, 1961.
- [22] R. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inf. Theory*, vol. 19, no. 2, pp. 244–246, March 1973.

- [23] A. G. D'yachkov, "Bounds on the average error probability for a code ensemble with fixed composition," *Prob. Inf. Transm.*, vol. 16, no. 4, pp. 3–8, 1980.
- [24] G. Poltyrev, "Random coding bounds for discrete memoryless channels," *Prob. Inf. Transm.*, vol. 18, no. 1, pp. 9–21, 1982.
- [25] J. Löfberg, "YALMIP : A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, Taipei, Taiwan, 2004.
- [26] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3rd Prague Conf. on Inf. Theory*, 1962, pp. 689–723, English Translation: <http://www.math.wustl.edu/~luthy/strassen.pdf>.
- [27] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.
- [28] T. S. Han, *Information-spectrum methods in information theory*. Springer, 2002.
- [29] D. Wang, A. Ingber, and Y. Kochman, "The dispersion of joint source-channel coding," <http://arxiv.org/abs/1109.6310>.
- [30] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Cost-constrained random coding and applications," in *Inf. Theory and Apps. Workshop*, San Diego, CA, Feb. 2013.
- [31] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. American Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
- [32] J. L. Jensen, *Saddlepoint Approximations*. Oxford University Press, 1995.
- [33] A. Martinez and A. Guillén i Fàbregas, "Random-coding bounds for threshold decoders: Error exponent and saddlepoint approximation," in *Int. Symp. Inf. Theory*, St. Petersburg, Aug. 2011.
- [34] Y. Altug and A. B. Wagner, "A refinement of the random coding bound," in *Allerton Conf. on Comm., Control and Comp.*, Monticello, IL, Oct. 2012.
- [35] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci.*, vol. 39, pp. 42–47, 1953.
- [36] R. Bahadur and R. Ranga Rao, "On deviations of the sample mean," *The Annals of Mathematical Statistics*, vol. 31, pp. 1015–1027, Dec. 1960.